Singlish Message Paraphrasing: A Joint Task of Creole Translation and Text Normalization

Zhengyuan Liu[†], **Shikang Ni**[‡,*], **Ai Ti Aw**[†], **Nancy F. Chen**[†]
[†]Institute for Infocomm Research, A*STAR, Singapore
[‡]University of Cambridge, UK
{liu_zhengyuan,nfychen}@i2r.a-star.edu.sg

Abstract

Within the natural language processing community, English is by far the most resource-rich language. There is emerging interest in conducting translation via computational approaches to conform its dialects or creole languages back to standard English. This computational approach paves the way to leverage generic English language backbones, which are beneficial for various downstream tasks. However, in practical online communication scenarios, the use of language varieties is often accompanied by noisy user-generated content, making this translation task more challenging. In this work, we introduce a joint paraphrasing task of creole translation and text normalization of Singlish messages, which can shed light on how to process other language varieties and dialects. We formulate the task in three different linguistic dimensions: lexical level normalization, syntactic level editing, and semantic level rewriting. We build an annotated dataset of Singlish-to-Standard English messages, and report performance on a perturbation-resilient sequence-to-sequence model. Experimental results show that the model produces reasonable generation results, and can improve the performance of downstream tasks like stance detection.

1 Introduction

While the development of natural language processing (NLP) has been focused on major languages such as English, Chinese, and French, there is emerging research interest in similar languages, varieties, and dialects (Zampieri et al., 2020). The distinction of language variations generally comes from geographical, historic, communicative settings, and social group dimensions (Coseriu, 1981). In particular, creole languages are formed in conditions when major languages (e.g., English, French) are adopted in another culture or region, and they often mix with existing languages and evolve into other varieties in their own right. Such examples include the French-based Haitian (Degraff, 1992), English-based Australian Kriol (Harris et al., 1993), and Colloquial Singaporean English (Singlish) (Ho et al., 1993). To adopt computational NLP solutions on dialects or creoles, applying existing models trained with major language resources will result in degraded performance, and collecting and annotating sufficient data for task-specified domain adaptation is time-consuming and labor-intensive. Similar to studies on multilingual scenarios (Balahur and Turchi, 2012; Eriguchi et al., 2018), one straightforward and effective approach is to conform the varieties to their base languages by machine translation (Zbib et al., 2012), then other NLP systems (e.g., sentiment analysis, information retrieval, reading comprehension) that take base languages as input could be applied. However, it would be challenging to apply this approach to dialects or creoles that include certain deviated grammar and local vocabularies with systems trained on the corresponding standard languages, especially to the under-resourced language varieties where machine translation performance is subpar (Nguyen and Chiang, 2017; Honnet et al., 2018).

Singlish, namely the colloquial Singapore English, is used in the daily lives of Singaporeans (Ho et al., 1993). Despite the obvious attribute of inheriting a large vocabulary base and foundational grammatical rules from English, Singlish imports terms and features from regional dialects including Mandarin, Malay, Hokkien, Teochew, Cantonese, and Tamil (Deterding, 2007), making its lexicon, syntax, and semantics deviate significantly from English (Wee, 2008; Wang et al., 2017). Thus, Singlish is an expressive language studded with colorful multicultural slangs yet manifested in an extremely condensed form (Leimgruber, 2011). As a result, a person only familiar with American or

[*] The contribution of Shikang Ni is conducted during his student internship in Institute for Infocomm Research, A*STAR, Singapore.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3924–3936
October 12–17, 2022.

Singlish Sentence Hhh. I kaypoh meh. I thought both of us also e kaypoh type. Like dat i dun ask lor...
Converted English Sentence Haha. Am I a busybody? I thought both of us are also the busybody type. Like that then I don't ask...
Singlish Sentence I thk boh bian one. What other topics you want them to talk to u abt ? metaverse meh ?
Converted English Sentence I think it is unavoidable. What other topics do you want them to talk to you about? like Metaverse?

Table 1: Two examples of the Singlish-to-English paraphrasing. Text in blue and red are involved with lexical normalization and creole translation, respectively.

British English, might have a difficult time understanding Singlish (see examples shown in Table 1). This also holds true for computational approaches, where mainstream and popular language models cannot be directly applied to Singlish.

Moreover, in practical use cases (e.g., social platform communication, SMS messages), the challenge is further complicated caused by various noise types in user-generated content, such as typos, spelling variations, phonetic substitutions, and ad hoc abbreviations (Sproat et al., 2001) (see examples shown in Table 1). This would further complicate the dialect or creole translation tasks. While various statistical and neural-based models are proposed for content de-noising in the form of text normalization (Supranovich and Patsepnia, 2015; Muller et al., 2019), much prior work only conduct word-level correction (Baldwin et al., 2015; van der Goot et al., 2021), and non-canonical English varieties are less studied.

Therefore, in this paper, we introduce a joint task of creole language translation and text normalization of Singlish messages. Since the deviations of Singlish from English come from both the lexical and the grammatical levels (Leimgruber, 2011), the task is conducted in a sentence paraphrasing manner. Based on the linguistic characteristics of Singlish, and user behavior of online communication, we further categorize the paraphrasing into three sub-tasks: lexical level normalization, syntactic level editing, and semantic level rewriting. Guided by this linguistic hierarchy, we build a dataset of Singlish-to-English paired messages annotated by human linguistic experts. We then evaluate a neural sequence-to-sequence approach on the paraphrasing task by fine-tuning state-of-the-art language backbones, and further optimize it with linguistic-featured input perturbation. We empirically show that the model can produce reasonable results, and downstream tasks can benefit from such text paraphrasing.[1] While our work focuses on Singlish, a special variant of English, the paraphrasing task formulation, linguistic analysis, and annotation protocol are general and can be extended to studying and processing other creole languages and dialects.

2 Related Work

NLP for Similar Languages Language variation (e.g., different dialects or national varieties of the same language) poses challenges for NLP applications, such as machine comprehension and dialogue systems. As a result, there is much of recent interest in computational processing of creoles and dialects (Zampieri et al., 2020). Related research areas include **language and dialect identification** (Suzuki et al., 2002; Lui et al., 2014; Zampieri et al., 2019) and **machine translation** (Altintas and Cicekli, 2002; Wang et al., 2016). Examples of machine translation between different dialects of the same language include British-American English (Zhao et al., 2000), Cantonese-Mandarin Chinese (Zhang, 1998), and European–Brazilian Portuguese (Costa-jussà et al., 2018). For closely related languages and dialects, many differences occur at the morphological level, thus word-for-word mapping, manual language-specific rules, and phrase-based statistical systems were proposed and applied (Hajič et al., 2000; Nakov and Tiedemann, 2012; Aharoni et al., 2019). Recently, data-driven neural approaches yield further improvement (Costa-jussà et al., 2018), and show the potential of transfer learning from one language pair to another (Nguyen and Chiang, 2017).

Text Normalization Online user-generated content is a valuable NLP resource, but it is often noisy and non-canonical. Most existing models are developed on canonical languages. Such models do not cope well with the disfluencies and informal phenomena (Karpukhin et al., 2019). Text normalization converts such noisy input to a 'normal' format (Sproat et al., 2001), while preserving the original meaning. Eisenstein (2013) studied several underlying factors that cause non-standard text like illiteracy and pragmatics. Since noise often comes

[1] Interested readers can contact corresponding authors for the data and code.

Level 1: Lexical Level Normalization
Lexical Variations in User-Generated Content Tackling the common user-generated lexical variations, including lower-case and upper-case (E.g., 'mrt' → 'MRT', 'TOdAy' → 'TODAY'), spelling typo (e.g., 'domian' → 'domain', 'r0bust' → 'robust'), single-word abbreviations (e.g., 'pple' → 'people', 'coz' → 'because'), phonetic substitutions (e.g., 'tym' → 'time', '4U' → 'for you'), and other non-standard spellings (e.g., 'goooood' → 'good'). **Lexical Variations in Singlish** Tackling the Singlish lexical variations, such as special short forms (e.g., 'yck' → 'YCK (Yio Chu Kang)'), and colloquial words (e.g.,'cheapo' → 'cheapskate', 'gahmen' → 'government'). **Non-English Word Borrowing** Replacing the non-English words borrowed from other languages that have a word-to-word mapping. E.g., 'mei mei' → 'sister' (Mandarin), 'pa tuo' → 'dating' (Cantonese), 'ta pau' → 'take-away' (Cantonese), 'huat' → 'to prosper' (Hokkien), and 'makan' → 'food' (Malay).
Level 2: Syntactic Level Editing
Missing Pronoun & Copula Recovering the appropriate pronouns, and the necessary verbs (e.g. "m typing a sms" → "I am typing a SMS", "oh cat so cute" → "oh the cat is so cute"). **Non-Standard Syntax & Grammar** Fixing the non-standard grammar in colloquial Singlish sentences, such as the topic prominence phenomenon (e.g., "A bit late lah, I came there." → "I came there a bit late."). **Missing Punctuation** Inserting the punctuation to where it is necessary (e.g., "Is that your book" → "Is that your book?").
Level 3: Semantic Level Rewriting
Colloquial Wording Some wording is different from colloquial Singlish and English, thus it needs to paraphrase the sentence while retaining the same semantic meaning (e.g. "Call aint going." → "The call is not coming through.") **Discourse Particles** Some clausal-final discourse particles indicate much semantic information (e.g., 'leh' marks a tentative request, 'lah' is a mood marker, and appeals for accommodation). For instance, "U leh, i going back liao." → "What about you? I am going back." **Non-English Spans & Code-Switching** Some non-English spans and the code-switching require clause or sentence level translation (e.g. "You sian? Let's go shopping!" → "Are you feeling bored? Let's go shopping!", "makan where?" → "where should we eat?").

Table 2: Three sub-tasks of the Singlish message paraphrasing.

from character/token level manipulation, early studies utilized lexical-based methods like dictionary lookup, word similarity, and N-gram probabilities (Han and Baldwin, 2011; Supranovich and Patsepnia, 2015). MoNoise (van der Goot and van Noord, 2017) built a pipeline that is similar to a ranking-retrieval approach. Recently, Muller et al. (2019) enhanced the BERT (Devlin et al., 2019) architecture so that the language model is able to add/remove tokens for word correction, and Bucur et al. (2021) applied a pre-trained language model for multilingual lexical normalization.

3 Singlish Message Paraphrasing Corpus

While unsupervised models show impressive results on tasks like semantic similarity matching by leveraging feature-rich pre-trained backbones (Devlin et al., 2019), their performance on language generation is still subpar. To foster data-driven approaches via supervised learning, we construct a human-annotated corpus for the Singlish message paraphrasing. The raw Singlish messages are extracted from the NUS Short Message Service (SMS) Corpus (Chen and Kan, 2013), which contains a total of 56K message samples from real-world mobile chats. We choose this resource since their data collection process employs Singaporean participants, and the SMS conversations cover a wide range of topics.[2] The annotation target is to convert the messages from colloquial Singapore English to standard American English, and the human-annotated references are expected to be understandable to non-Singaporean high school students.

3.1 Paraphrasing Sub-task Definition

Online communication between creole users is often a mix of language-specific usage and noisy content generation. The paraphrasing task of Singlish

[2]All samples we use are from the published anonymized dataset, and do not contain any personal information.

messages thus requires text editing from multiple aspects. Combining our analysis of real-word Singlish messages, and previous linguistics and lexical normalization studies (Wee, 2008; van der Goot and van Noord, 2017), we categorize this paraphrasing task into three sub-tasks: lexical level normalization, syntactic level editing, and semantic level rewriting.[3]

Lexical Level Normalization Lexical normalization is to uniform the non-standard tokens and borrowed words via infilling and replacement. We first tackle the common English lexical variations like typos, abbreviations, phonetic substitutions, and misspellings which are ubiquitous in online communication platforms (Sproat et al., 2001; Supranovich and Patsepnia, 2015). Moreover, in Singlish messages and conversations, there are special short forms, discourse particles, and words borrowed from other languages (e.g., Mandarin, Malay) (Leimgruber, 2011). In addition to an existing localized vocabulary,[4] the annotators were asked to collect a list of such special words, and some examples are shown in Table 2.

While standard word recovering often can be done independently without sentence understanding, in some cases, the context is necessary for disambiguation. For instance, the word *'goooood'* in *"U r so goooood."* and *"really? oh my goooood!"* should be converted to *'good'* and *'god'*, respectively. Another case is converting some phonetic substitutions, for example: *"This is a gift 4U!"* → *"This is a gift for you!"*. In addition, it is difficult to obtain a complete collection of all non-standard tokens and borrowed words. Therefore, non-computational vocabulary-based methods are not sufficient for lexical normalization (Muller et al., 2019; Bucur et al., 2021), and in our setting, it becomes one sub-task of the sequence-to-sequence modeling.

Syntactic Level Editing The grammar of Singapore English differs from the standard English markedly. For example, some pronouns and BE verbs are often omitted (e.g., *"the weather hot lah"*). Another language-specific phenomenon is the feature 'topic prominence', where the topic span (e.g., noun phrases) is re-ordered to the beginning of the sentence (e.g., *"this book last year i read"* → *"I read this book last year."*). This construction in colloquial Singlish is adapted from Chinese and Malay (Leimgruber, 2011), and the topic prominence can be further highlighted by the insertion of a break or a discourse particle between the topic and the clause (e.g *"Too slow (lah), I find that building"*). Moreover, in online communications, users tend to omit punctuation, especially at the end of sentences (e.g., question marks). Thus, we also take punctuation into consideration in the annotation protocol.

For the sub-task syntactic level editing, sentences are converted to a standard American English grammar, which often requires changes of more than one word or span.

Discourse Particle: 'leh' marks a tentative suggestion or request.
Original Text: Still eating. Got free mcflurry. U (leh), going back liao.. **Paraphrased Text:** I am still eating. I got a free McFlurry. What about you? I am going back.
Discourse Particle: 'hor' attempts to garner support for a proposition. **Original Text:** I go can (hor).. **Paraphrased Text:** Is it alright for me to go?

Table 3: Two examples of clause-final discourse particles in colloquial Singlish.

Semantic Level Rewriting As the narrative and wording of the same meaning are different from Singlish and English, sometimes it needs rewriting the sentence while retaining the same semantic meaning. Particularly, the usage of clausal-final discourse particles (e.g., *'leh'*, *'hor'*), which originates from Hokkien and Cantonese, is one of the most well-known features of Singlish, and some fillers indicate much semantic information (Leimgruber, 2011). For instance, as shown in Table 3, the discourse particle *'hor'* conveys inquiring meaning, thus the sentence *"I can go (hor)"* should be converted to *"Is it alright for me to go?"*. More discourse particles and their examples are shown in Appendix Table 10.

Moreover, in real-world online communication and inter-language scenarios, some non-English spans (involving Singlish and cross-language usage where the syntax is mostly on one language) and the code-switching phenomenon require clause or sentence level translation and rephrasing instead of word level replacement. For instance, the sentence

[3]While colloquial Singapore English presents various lexical and grammar features, here we focus more on those which significantly affect language understanding. Features like tense agreement (Leimgruber, 2011) are not considered if the context is insufficient, to reduce annotation variance.

[4]http://www.singlishdictionary.com/

"makan where?" (the word *'makan'* in Malay means *'eat'*) should be converted to *"where should we eat?"*. This is also included in the semantic level rewriting sub-task.

3.2 Corpus Construction

3.2.1 Sample Pre-processing

We draw our source material from the texts of the NUS Short Message Service (SMS) Corpus (Chen and Kan, 2013), which consists of 56K messages originating from Singaporeans and university students. In our data pre-processing, we first filtered out raw samples that are shorter than 20 characters, as well as duplicated items, resulting in a 45K data size. We then sampled the representative Singlish messages, to refine the subset for annotation. Here the sentence-level perplexity value calculated with a language model GPT-2 (Radford et al., 2019) is used as the criteria, where a high perplexity score indicates the text is more distinct than the canonical English.[5] We then ranked all samples accordingly, and kept those above average.

3.2.2 Data Annotation

The annotation of paraphrasing task is conducted with a group of 6 linguistic experts from a local university, who are proficient in both Singlish and English. To facilitate the process and reduce the annotation variance across different annotators, they are asked to complete the three sub-tasks hierarchically from the low level (lexical normalization) to the high level (semantic rewriting) with minimum changes. Moreover, as the SMS corpus retains the original order of messages in a conversation, we did not shuffle them, thus annotators can refer to the context for better paraphrasing.

Since it is challenging to perfectly paraphrase all Singlish messages, we allow participants to assign confidence scores to their annotations, which indicate the level of agreement between two of any annotators on the same sample, according to the sus-tasks defined in Section 3.1. Confidence scores are from low (0) to high (5) agreement level, and the score of 0 is labeled when the necessary context for rephrasing is missing or the whole source sentence is written in a non-English language (e.g., Tamil, Malay), and such samples are excluded in the training set.

[5]https://huggingface.co/docs/transformers/perplexity

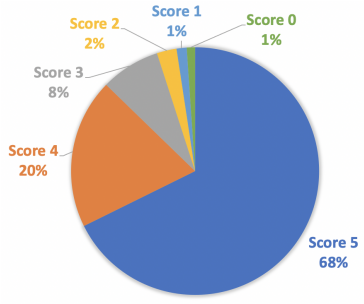

Figure 1: Distribution of confidence scores on the annotated samples. Samples with score 0 are considered as invalid annotation, and excluded from the dataset.

3.2.3 Annotation Analysis

The confidence score distribution of our annotated set is shown in Figure 1. We observed that the scores under average usually result from the semantic level rewriting, especially for non-English span translation. Moreover, to assess the text level variance among annotators, 1500 samples across all confidence levels are randomly selected and annotated by two annotators. Following the common automatic evaluation in machine translation (Aharoni et al., 2019), we use BLEU (Papineni et al., 2002) as the metric, and the average score calculated on those samples is 69.7, which shows a reasonable human annotation agreement.

4 Automatic Paraphrasing via Sequence-to-Sequence Modeling

4.1 Base Neural Architecture

Unlike previous lexical normalization work (Aw et al., 2006; Baldwin et al., 2015), the syntactic and semantic level sub-tasks in our setting require a higher capability of contextual understanding and sentence generation. Therefore, we introduce sequence-to-sequence modeling for Singlish message paraphrasing. Define x as the input text, and y as the target output. A neural encoding-decoding model G is used. The goal is formulated to maximize $P(y|x;\theta_G)$, where θ_G are the learnable parameters. In our settings, the base architecture is a Transformer-based auto-regressive language model, since the Transformer (Vaswani et al., 2017) shows strong capabilities of contextual modeling and generation, and is widely adopted in various natural language processing tasks (Radford et al., 2019; Lewis et al., 2020). The encoder consists of a stack of Transformer layers. Each layer has two sub-components: a multi-head layer with self-attention

mechanism, and a position-wise feed-forward layer. A residual connection is employed between each pair of the two sub-components, followed by layer normalization. The decoder also consists of a stack of Transformer layers. In addition to the two sub-components in the encoding layers, the decoder inserts another component that performs multi-head attention over hidden representations from the last encoding layer. Then, the decoder generates tokens from left to right in an auto-regressive manner. The architecture and formula details are described in (Vaswani et al., 2017).

With the parallel message pairs, we conduct supervised learning with token-level maximum likelihood estimation. At the training stage, the cross-entropy loss is calculated between the decoder's output and the reference sentence:

$$l(\theta) = -\Sigma_i \log(p(y_i|y_{1:i-1}, x; \theta_G)) \quad (1)$$

4.2 Linguistic-featured Input Perturbation

While fine-tuning language backbones bring about impressive performance on cross-language translation (Liu et al., 2020), they are vulnerable to noisy input. Moreover, data-driven approaches may overfit to superficial lexical features rather than learn how to paraphrase from a semantic aspect, especially when the training data are limited.

To tackle these two challenges, we adopt a simple yet effective model enhancement via input perturbation, inspired by the linguistic characteristics of the Singlish messages, and the denoising sequence-to-sequence pre-training scheme (Lewis et al., 2020). There are two operations: (1) Word Perturbation: To simulate lexical variations in real-world online user-generated content, we randomly remove or replace one character of each word (with a 10% probability). Here the word perturbation is only conducted on words from a common English word vocabulary,[6] excluding terminology words, named entities, fillers, and special short forms. Moreover, recent studies show that character-level noise makes models more robust towards spelling variations (Aepli and Sennrich, 2022). (2) Sentence Perturbation: To simulate the grammatical features like topic prominence, and enhance contextual modeling of the language backbone, we randomly inject noise by exchanging a bi-gram pair in each sentence with a 10% probability.

[6]https://github.com/first20hours/google-10000-english

Corpus	Size	Task Type
English Tweets (Baldwin et al., 2015)	3K	Lexical
English Message (Aw et al., 2006)	5K	Lexical
Our Singlish SMS Corpus	20K	All levels

Table 4: Statistics of the corpora used in our setting. The English tweets and message corpora are only for lexical normalization, and are used at the warm-up training stage. See corpus combination results in Table 9.

5 Experiments

5.1 Training Datasets

The Singlish message corpus built in Section 3 is used for model training and evaluation. The training, validation, and test set size are 20000, 1000, and 1000, respectively. In addition to the Singlish dataset, we include two English lexical normalization corpora (Aw et al., 2006; Baldwin et al., 2015) in the warm-up training stage. Data statistics are shown in Table 4. For samples we annotated, the average sentence number per message is 2.13, and the average word number is 18.85.[7] While these two datasets are much smaller and only focus on lexical normalization, in our pilot experiments, we empirically observed that warm-up training with the additional data brings 2-3% relative improvement consistently (see Table 9).

5.2 Experiment Setup

In our experimental setting, we first trained and evaluated the base model: a Vanilla Transformer (Vaswani et al., 2017) (6 encoder and 6 decoder layers, with 768 hidden size and a fixed token embedding layer). We then incorporated prior language knowledge to the base model by loading the *BART-base*, *BART-large*, and *mBART* (Lewis et al., 2020; Liu et al., 2020), and fine-tuned the backbones with parallel pairs. For automatic evaluation metrics, we adopted the common methods used for language generation based on n-gram overlap: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and the semantic-based metric BERTScore (Zhang et al., 2020) upon similarity of contextualized sentence representations.

All models were implemented with PyTorch and Hugging Face Transformers [8]. AdamW optimizer (Kingma and Ba, 2015) was used. The batch size and learning rates were set at 16 and 2e-5, respec-

[7]https://www.nltk.org/api/nltk.tokenize.html
[8]https://github.com/huggingface/transformers

Model Type	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Unaltered Singlish Messages (Lower bound)	35.6	61.1	63.7	42.5	68.7	72.2
Vanilla Transformer (Vaswani et al., 2017)	47.7	72.3	75.4	65.3	74.1	78.5
BART-base w/ Fine Tuning	55.0	79.7	83.5	70.2	85.7	84.1
BART-base w/ Fine Tuning + Input Perturbation	57.8	81.1	84.6	72.3	86.1	86.5
BART-large w/ Fine Tuning	58.4	81.0	85.0	73.5	86.3	86.4
BART-large w/ Fine Tuning + Input Perturbation	**61.5**	**81.6**	**85.8**	**75.8**	**87.0**	**87.9**
mBART-large-50 w/ Fine Tuning	57.2	80.3	84.6	72.3	85.3	86.0
mBART-large-50 w/ Fine Tuning + Input Perturbation	60.2	81.1	85.2	74.9	85.9	86.7

Table 5: Automatic evaluation scores on the Singlish message paraphrasing task. The proposed text perturbation is conducted on all input samples at the training stage. ROUGE and BERTScore reported here are F1 scores.

tively. We added label smoothing (weight λ = 0.1) on the cross-entropy loss (Müller et al., 2019). Warm-up training step number was 2000. We used early stopping (patience = 5) if validation performance did not improve. Test results were reported with the best validation checkpoints. Beam search size was set at 5. Other information such as environment details and trainable parameter sizes are shown in Appendix Table 12.

5.3 Automatic Evaluation Results

We first calculate the evaluation metrics between **Unaltered Singlish Messages** and annotated text. This serves as a lower bound performance, as no paraphrasing is conducted (see Table 5), which also demonstrates Singlish bears unique usages from standard English. Compared to the **Vanilla Transformer**, leveraging pre-trained language backbones significantly improves the performance, and the **BART-large** outperforms the **BART-base**. Adopting the input perturbation further yields certain improvements. To evaluate the effectiveness of leveraging a multilingual backbone, we also applied **multilingual BART** (mBART) (Liu et al., 2020). However, it did not show any additional performance gains. Presumably, this is because borrowing features and non-English words/spans in Singlish are not well represented in the multilingual pre-training process and data resources. For example, Hokkien and Malay are not in the supported language list of mBART. Moreover, in Singlish messages, the Mandarin words are not expressed in Chinese characters but in Pinyin, which are currently not included in the pre-training of most multilingual backbones.

5.4 Human Evaluation Results

Aside from automatic evaluation, we conducted a human evaluation to complement objective metrics. Following prior work (Wieting and Gimpel, 2018), each text candidate is scored on a five-grade scale of [1, 5], where 1 means the paraphrasing is unacceptable, and 5 means it can be taken as a ground truth. We randomly selected 100 test samples, and asked the linguistic experts to score the corresponding human-written and model-generated outputs. Details of the assessment interface are shown in Appendix Figure 3. Six raters conducted the human evaluation independently, and the average scores are summarized and shown in Table 6. While input perturbation brings significant improvement, there is still space for models to reach human reference performance.

Model Type	Avg. Rating Score
Human Reference	3.87
BART-large w/ Fine Tuning	3.12
+ additional Input Perturbation	3.41

Table 6: Human evaluation results. 100 samples were randomly selected from the test set and assessed by 6 linguistic experts. All rating scores are averaged.

To gain further insights into the limitations of automatic paraphrasing, we conduct text-level analysis on some samples. As shown in Table 7, neural generators produce reasonable changes in the lexical normalization and syntactic editing, while semantic rewriting is still relatively challenging, especially for code-switching and some non-English spans (see more examples in Appendix Table 11). Considering the insufficient language modeling of borrowing words and special abbreviations and the limited corpus with human annotation, we speculate an augmented unsupervised pre-training process is beneficial for tackling this challenge, and it can be one of the future work.

5.5 Experiment on Different Sample Groups

The confidence scores described in Section 3.2.2 are labeled during the annotation process. They can present the inter-annotator agreement level, and partially reflect the sample difficulty of paraphrasing.

Model Type	Text Content
Source Input	Yup... Okay. Cya tmr.. So long nvr write already... Dunno whether tmr can come up with 500 words
Human Reference	Yes... Okay. See you tomorrow... It has been so long since I have written... I do not know whether tomorrow I can come up with 500 words.
BART-large w/ Fine Tuning	Yes... Okay. See you tomorrow... So long never write already... Do not know whether tomorrow I can come up with 500 words.
BART-large w/ Input Perturbation	Yes... Okay. See you tomorrow... I have not written in so long already... I do not know whether tomorrow I can come up with 500 words.

Table 7: One Singlish message example and the generated text from human annotation and neural models. Text spans colored in blue are appropriate changes, and in purple are sub-optimal changes.

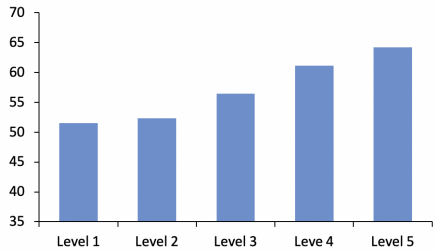

Figure 2: BLEU scores calculated on test sample groups with different confidence scores (range from 1 to 5). Samples with score 0 are considered as invalid annotation, and excluded from the test set.

Model Type	Precision	Recall	F1 Score
Training and Evaluation on Original Samples			
BERTweet-base	68.6	72.1	70.0
RoBERTa-base	70.8	71.7	71.1
Training and Evaluation on Processed Samples			
BERTweet-base	70.4	72.5	**71.2** [1.7% ↑]
RoBERTa-base	71.3	74.3	**72.7** [2.3% ↑]

Table 8: Results on the stance detection task. We processed the corpus with our message paraphrasing model for comparison to raw samples. Values in bracket denote the relative performance increase.

Therefore, we calculate BLEU scores on test sample groups with different confidence scores. As shown in Figure 2, the BLEU scores of level 4 and 5 are larger than those of level 1 and 2, and this demonstrates that samples with lower confidence scores are generally more challenging for both human and automatic paraphrasing.

5.6 Experiment on Corpus Combination

To assess the effectiveness of joint training of lexical normalization and sentence paraphrasing as well as their difference, we further conduct an experiment upon the single and mixed training data combination, as shown in Table 9. From the result, we observed that: (1) models only trained on lexical normalization corpora could not provide strong baseline performance on our Singlish paraphrasing task. (2) compared with single training on our Singlish corpus, training on the merged dataset yields 2-3% relative improvement at all fronts.

5.7 Experiment on Downstream Task

When applied to online communication, the paraphrasing model is able to reduce text noise such as non-canonical wording and misspelling, and it is potentially beneficial for various downstream tasks where noisy samples are ubiquitous. In this article, we choose stance detection of English tweets (Mohammad et al., 2016) for experimentation, which is generally formulated as a classification problem of 3 types (i.e. *Favor*, *Against*, and *None*). The corpus consists English samples with non-standard lexical and syntactic features. We ran the paraphrasing model (*BART-large w/ Fine Tuning + Input Perturbation*) on both training and test tweet samples. Then following previous work, two strong and representative baselines *BERTweet* (Nguyen et al., 2020) and *RoBERTa* (Liu et al., 2019) are trained on the processed corpus (more configuration details are shown in Appendix Table 12), and we reported F1, precision and recall scores on the test set. As shown in Table 8, while the two models perform slightly differently, their classification performance obtained improvement on all fronts after the de-noising transformation (especially a 2.3% relative F1 score increase). This suggests that while our paraphrasing model is trained on Singlish messages, it is still useful for tasks that are not in Singlish since it learned re-writing from context and can reduce the input noise significantly.

6 Conclusions

In this paper, we analyzed the representative linguistic features of colloquial Singapore English, and proposed a joint task of creole language translation and text normalization. We formulated the

Model Type	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Unaltered Singlish Messages (Lower bound)	35.6	61.1	63.7	42.5	68.7	72.2
Training on two English lexical normalization corpora						
BART-base w/ Fine Tuning	45.9	65.7	77.7	63.1	80.2	77.1
BART-base w/ Fine Tuning + Input Perturbation	47.0	67.3	79.1	64.8	82.2	78.8
Training on our Singlish message corpus						
BART-base w/ Fine Tuning	53.3	78.9	81.8	69.2	83.9	80.3
BART-base w/ Fine Tuning + Input Perturbation	56.3	80.1	83.0	71.4	84.5	84.8
Training on the merged dataset						
BART-base w/ Fine Tuning	55.0	79.7	83.5	70.2	85.7	84.1
BART-base w/ Fine Tuning + Input Perturbation	57.8	81.1	84.6	72.3	86.1	86.5

Table 9: Corpus combination experiment with automatic evaluation scores. Models are trained separately on English lexical normalization and our Singlish message data. The proposed text perturbation is conducted on all input samples at the training stage. ROUGE and BERTScore reported here are F1 scores.

paraphrasing task of Singlish-to-standard English into three sub-tasks: lexical level normalization, syntactic level editing, and semantic level rewriting. Based on this linguistic hierarchy, we constructed an annotated dataset and reported baseline performance via fine-tuning language backbones, and further robustified the neural models with linguistically-inspired input perturbation. Experiment on a downstream stance detection task showed better performance when the input (colloquial English or Singlish) is de-noised by our paraphrasing model, suggesting that models developed using the data we build could help normalize noisy user-generated text. Our task definition, annotation protocol, constructed corpus, and reported base results pave the way for future studies on creole and colloquial language processing.

Acknowledgments

This research was supported by funding from the Institute for Infocomm Research (I2R) under A*STAR ARES, Singapore. We thank Hai Leong Chieu for the insightful discussions, and Jia Yi Chan, Nabilah Binte Md Johan, Siti Maryam Binte Ahmad Subaidi, and Siti Umairah Md Salleh for linguistic resource construction, and human evaluation. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3874–3884.

Kemal Altintas and Ilyas Cicekli. 2002. A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*, pages 192–196. Citeseer.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40.

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52–60.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

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.

Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. 2021. Sequence-to-sequence lexical normalization with multilingual transformers. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 473–482.

Tao Chen and Min-Yen Kan. 2013. Creating a live, public short message service corpus: the nus sms corpus. *Language Resources and Evaluation*, 47(2):299–335.

Eugenio Coseriu. 1981. Los conceptos de dialecto, nivel y estilo de lengua y el sentido propio de la dialectología. *LEA: Lingüística española actual*, 3(1):1–32.

Marta R Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282.

Michel Anne Frederic Degraff. 1992. *Creole grammars and acquisition of syntax: The case of Haitian*. Ph.D. thesis, University of Pennsylvania.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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,*, pages 4171–4186.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proceedings of the sixth conference on Applied natural language processing*, pages 7–12.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 368–378.

John Harris et al. 1993. Losing and gaining a language: the story of kriol in the northern territory. *Language and culture in Aboriginal Australia*, page 145.

Mian Lian Ho, John Talbot Platt, et al. 1993. *Dynamics of a contact continuum: Singaporean English*. Clarendon Press.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing swiss german. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. *W-NUT 2019*, page 42.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the ACL 2020*, pages 7871–7880.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952.

Benjamin Muller, Benoît Sagot, and Djamé Seddah. 2019. Enhancing bert for lexical normalization. In *The 5th Workshop on Noisy User-generated Text (W-NUT)*.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When does label smoothing help? *arXiv preprint arXiv:1906.02629*.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305.

Dat Quoc Nguyen, Thanh Vu, and Anh-Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *IJCNLP 2017*, page 296.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer speech & language*, 15(3):287–333.

Dmitry Supranovich and Viachaslau Patsepnia. 2015. Ihs_rd: Lexical normalization for english tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 78–81.

Izumi Suzuki, Yoshiki Mikami, Ario Ohsato, and Yoshihide Chubachi. 2002. A language and character set determination method based on n-gram statistics. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(3):269–278.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Rob van der Goot and Gerardus van Noord. 2017. Monoise: Modeling noise using a modular normalization system. *Computational Linguistics in the Netherlands Journal*, 7:129–144.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *In Advances in Neural Information Processing Systems*, pages 6000–6010.

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.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics*, 42(2):277–306.

Lionel Wee. 2008. *Singapore English: morphology and syntax*, pages 2250–2264. De Gruyter Mouton.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, et al. 2019. A report on the third vardial evaluation campaign. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 2020*.

Xiaoheng Zhang. 1998. Dialect mt: a case study between cantonese and mandarin. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Liwei Zhao, Karin Kipper, William Schuler, Christian Vogler, Norman Badler, and Martha Palmer. 2000. A machine translation system from english to american sign language. In *Conference of the Association for Machine Translation in the Americas*, pages 54–67. Springer.

A Appendix

Discourse Particle	Original Text	Paraphrased Text
'leh' marks a tentative suggestion or request.	Still eating. Got free mcflurry. U (leh), going back liao..	I am still eating. I got a free McFlurry. What about you? I am going back.
'hor' attempts to garner support for a proposition.	I go can (hor)..	Is it alright for me to go?
'wot' marks obviousness and contradiction.	Datz (wot) i tld u 2 go 4 sleep...	That is why I told you to go to sleep...
'lor' indicates obviousness or resignation.	Yar (lor)... How u noe? U used dat route too?	Yes... How do you know? you used that route too?
'mah' marks information as obvious.	Lol that is my lunch timing (mah) wat time u want?	Lol that is my lunch time. What time do you want?
'meh' indicates skepticism.	s quite 4 me... U dun find it accurate (meh)...	It is quite for me... Don't you find it accurate?
'liao' present past tense.	They have done the test (liao)	They have already done the test.
'lah' is a mood marker, and appeals for accommodation.	She should be quite safe (lah)	She should be quite safe.
'one' indicates topic prominence (a syntax to emphasize the topic–comment structure).	This country weather very hot (one).	The weather is very warm in this country.

Table 10: Selected clause-final discourse particles of colloquial Singlish. In our annotation protocol, the particles that only convey (weak) emphasis will not trigger the semantic rewriting, such as *'ya'* and *'mah'*.

Model Type	Text Content
Source Input	Hehe cumi dnt mind much for litle quarels in ua lyf.. J0y n sadnes b0th are a part of lyf.. Try to smile in pain N atlast i'm wid u f0reva as a frnd..
Human Reference	Hehe Cumi did not mind much for the little quarrels in your life... Joy and sadness are part of life... Try to smile in pain and at least I am with you forever as your friend.
BART-large w/ Fine Tuning	Hehe cumi don't mind much for little quarels in your life.. Jokes and sadnes are a part of life.. Try to smile in pain and at last i'm with you f0reva as a friend.
BART-large w/ Input Perturbation	Hehe Cumi do not mind much for little quarels in your life.. Joy and sadness both are a part of life.. Try to smile in pain and at last I am with you forever as a friend..
Source Input	Woot.wan ask u ST de interview how de ah. will ask simi kind of thing.
Human Reference	Woot. I want to ask you how is the ST interview like. What kind of thing will they ask?
BART-large w/ Fine Tuning	Woot. Want to ask you ST the interview. Will ask simi kind of thing.
BART-large w/ Input Perturbation	Woot. I want to ask you ST how the interview. I will ask some kind of thing.
Source Input	Ahh I know lol. Like that if im the nominated one, walk pass also paiseh la
Human Reference	Ah, I know lol. If I am nominated, it will be embarrassing to walk past.
BART-large w/ Fine Tuning	I know lol. Like that if I am the nominated one, walk pass also sorry.
BART-large w/ Input Perturbation	Ahh, I know lol. Like that if I am the nominated one, walk pass also sorry.
Source Input	Haha. I was kidding lah. Yaloh. Yaloh. I'll try dessert also. Tml uintro their zao pai cai ar. Haha.
Human Reference	Haha. I was kidding. Yes. Yes. I will try dessert also. Tomorrow you will introduce their most famous dish. Haha.
BART-large w/ Fine Tuning	Haha. I was kidding. Yaloh. I will try dessert also. Tomorrow uintro their zao pai cai. Haha.
BART-large w/ Input Perturbation	Haha. I was kidding. Yes. I will try dessert also. Tomorrow I will sample their zao pai cai. Haha.
Source Input	baobei still syncing... dropbox so slow.. dar wait awhile morr
Human Reference	Baby, I am still syncing it... Dropbox is so slow... Dear, wait for a while.
BART-large w/ Fine Tuning	Baobei, it is still syncing...Dropbox is so low... Darling wait awhile more..
BART-large w/ Input Perturbation	Baby it is still syncing... Dropbox is so slow... darling, wait a while..

Table 11: Examples of Singlish message paraphrasing, and the generated text from human annotation and models. To improve the readability, here we only color the spans with sub-optimal changes in purple.

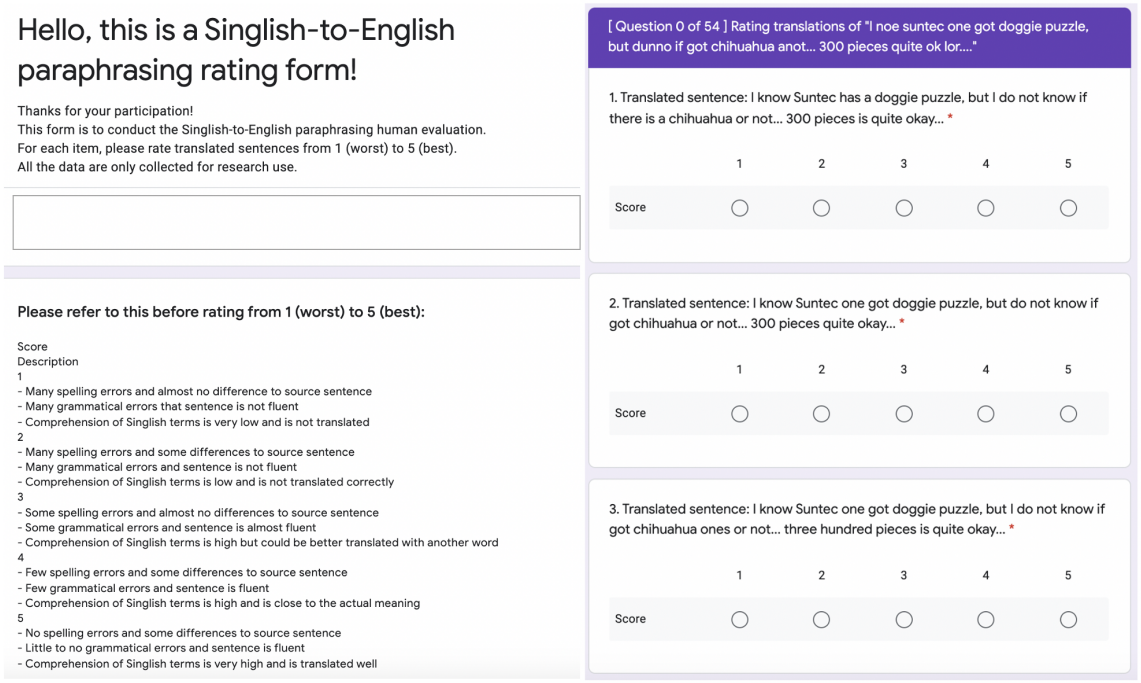

Figure 3: The rating form template for the human evaluation described in Section 5.4. Text candidates are shuffled for each sample to reduce the order bias, and we average the scores from all raters.

Environment Details	
GPU Model	Single Tesla V100 with 16 GB memory; CUDA version 10.1.
Library Version	Pytorch==1.7.1; Transformers==4.8.2.
Computational Cost	Average 1.5 hours training time for one round. Average 3 rounds for each reported result (calculating mean of the result scores).
Hyper-parameter	**Setting Detail**
Paraphrasing Task	
Neural Generator	Vanilla Transformer (12-layer, 768-hidden, 16-heads, 135M parameters). BART-base (12-layer, 768-hidden, 16-heads, 139M parameters). BART-large 24-layer, 1024-hidden, 16-heads, 406M parameters. mBART-large-50 (24-layer, 1024-hidden, 16-heads, 610M parameters).
Learning Rate and Batch Size	We set the learning rate (2e-5) and batch size (16) according to regular language model fine-tuning strategy (Lewis et al., 2020).
Beam Search Size	We evaluated beam search sizes from 3 to 10, and 5 provided the best balance of performance and inference speed.
Label Smoothing Weight	We set the label smoothing weight λ at 0.1 for fine-tuning on language generation work (Lewis et al., 2020).
BERTScore Metrics	We use the RoBERTa-base version of BERTScore (Zhang et al., 2020).
Stance Classification Task	
Corpus	The corpus we used for stance detection is from a published work (Mohammad et al., 2016), where all data are anonymized, and only for research use.
Neural Classifier	Bertweet-base (12-layer, 768-hidden, 12-heads, 130M parameters). RoBERTa-base (12-layer, 768-hidden, 12-heads, 125M parameters). Bertweet-base (12-layer, 768-hidden, 12-heads, 130M parameters).
Learning Rate and Batch Size	We set the learning rate (2e-5) and batch size (32) according to regular language model fine-tuning strategy (Devlin et al., 2019).

Table 12: Details of the experimental environment and the hyper-parameter setting.