Exploiting Available Resources for the Training of Manglish Language Models

Meisin Lee
Independent Researcher
meisin123@gmail.com

Lay-Ki Soon
Monash University (Malaysia Campus),
Jalan Lagoon Selatan, Bandar Sunway,
47500, Selangor, Malaysia
soon.layki@monash.edu

Abstract
Manglish (Malaysian Colloquial English) is a type of creole\(^1\) widely used in Malaysia. Like other creoles, Manglish is under-resourced and ignored in NLP literature. Manglish is made up of English as source language but also draws parts of its grammar and vocabulary from its parent languages such as Malay, Chinese (including mandarin and dialects) and Tamil. Due to the severe lack of resources, this work leverages on related-resources to train language models on Manglish. We look into two main categories of related resources: (1) available corpora and (2) existing pre-trained language models. For corpora, Singlish (Singaporean Colloquial English) and various code-switched corpora are used; as for pre-trained language models, existing models such as Singlish and other code-switched LMs involving its ‘parent languages’ are exploited. Given that Manglish is made up of resource-rich parent languages, i.e. English and Chinese, efficient processing of Manglish can allow us to tap into language technologies developed for these highly resourced languages.

1 Introduction
Manglish, or known as Malaysian Colloquial English\(^2\), is a type of creole language used widely in Malaysia. Although not an official written language, Manglish has become an important lingua franca in the multi-ethnic and multilingual community in Malaysia and is an important part of the local identity. Figure 1 shows an example of a typical Manglish sentence.

Manglish is considered an under-resourced language. Based on the best of our knowledge, there is no research done on Manglish as yet, nor is there any known Manglish corpus or any language model developed for it. Interestingly, the ‘parent languages’ of Manglish, namely English and Chinese, are two of the most highly-resourced languages while Bahasa Indonesia (a standardized variety of Malay) is one of the most spoken languages in the world. It is also highlighted in (Lent et al., 2021) that creole language at the sentence level, share commonalities with code-switched languages\(^3\). In the case of Manglish, can it be considered code-switching of a number of languages?

There is a lot of potential in Manglish research, where accurate processing can help us tap into and exploit the vast resources and language technologies developed for its parent languages. This paper sets out to collect existing resources related to Manglish to produce Language Models useful for other Manglish-related downstream tasks and subsequently, for the development of suitable language technologies.

2 Available Resources
Given that there is no known Manglish corpus, we look to existing related-resources. Manglish is very similar to and highly mutually intelligible with Singlish, a creole of similar roots. There is little distinction between the two creoles except that Manglish vocabulary contains more Malay words while Singlish contains more words of Hokkien (Min Nan) and Teochew origin, see Appendix A for

\(^{1}\)According to Wikipedia, a creole language is a stable natural language developed from a mixture of different languages.

\(^{2}\)or also known as Mangled English for the level of distortion done to standard English.


Figure 1: An example of a Manglish sentence. In standard English, it means “Hey, aren’t you hungry? I am eating at the nearby coffee shop.”
<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Name, Description</th>
</tr>
</thead>
</table>
| Creole - Singlish | National University Singapore SMS Corpus (Chen and Min-Yen, 2015)  
| Creole - Singlish | Universal Dependencies for Colloquial Singaporean English (Wang et al., 2017)  
| Creole - Singlish | webcrawler for Singaporean hardware forum (Tan et al., 2020)  
| Code-Switch     | A Mandarin-English Code-Switching Corpus: South East Asia Mandarin-English (SEAME) corpus (Lee et al., 2017)  
| Code-Switch     | Malay-English Code-Switching Subjectivity Corpus for Sentiment Analysis (Kasmuri and Basiron, 2019)  
| Language Model  | Singlish Language Model released by (Lent et al., 2021)  

Table 1: Available resources potentially useful for the building to a Manglish Language Model

examples. Besides this, we consider code-switch corpus involving Manglish’s parent languages, i.e., English-Chinese as well since it is common in Manglish to use sentence structures from Chinese. Table 1 captures the list of available resources for the building of a Manglish Language Model.

In terms of our own resources, we have collected some Manglish text from subreddits r/malaysia and lowyat.net, a local popular online forum. The corpus is about 1M lines of text with 15million word tokens, 125k word types.

3 Experiments and Research Questions

Experiments are designed to answer the following research questions:

- Can the existing Singlish language model be adapted to produce a Manglish language model?
- Given the slang and spelling variations of romanized chinese words, how should out-of-vocabulary words be dealt with? Does existing cross-lingual word embedding techniques such as multilingual meta-embeddings (MME)(Winata et al., 2019a) and hierarchical meta-embeddings (HME)(Winata et al., 2019b) work?
- How to identify and process intra-word code-switching, for example: ‘makan-ning’ in Figure 1 is made up of ‘makan’, a Malay word and the English suffix ‘ing’; it means ‘eating’.

4 Preliminary Experiments and Results

As part of the preliminary experiments, we used the Singlish model made available by (Lent et al., 2021) to further pre-train on Manglish text on the masked language modeling task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Prec</th>
<th>PrecD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singlish(Lent et al., 2021)</td>
<td>46.77</td>
<td>42.89</td>
</tr>
<tr>
<td>Manglish(further pre-training from the Singlish model)</td>
<td>35.56</td>
<td>45.54</td>
</tr>
</tbody>
</table>

Table 2: Results of Singlish and Manglish Language models based on masked language modeling. Results of Singlish is as per reported in (Lent et al., 2021). Prec is the precision score while PrecD is the precision using the authors’ own creole dictionary.

The above experiment aims to answer the first research question in Section 3. The results shown in Table 2 shows that the result for precision score for Manglish has deteriorated as compared to its parent model, Singlish. This is because even though Manglish shares a high degree of similarity with Singlish, there are special words unique to Manglish vocabulary. Further experiments will be conducted to answer the rest of the research questions.

5 Conclusion

Despite having 32 million speakers, Manglish is severely under-resourced and thus, is neglected in NLP research. We envisioned that language applications designed specifically for Manglish is of high utility value where it can acts as a bridge to tap into existing applications developed for highly resourced languages such as English and Chinese.
A Difference between Singlish and Manglish

The most significant difference between colloquial Malaysian English (Manglish) and colloquial Singaporean English (Singlish) is the different composition of Malay / Chinese and Chinese Dialect words. Manglish uses more Malay words and slangs while Singlish uses more Chinese words, due to their respective ethnic composition. This is illustrated with the examples below:

1. Singlish: The test was so hard, I just tikam only lah!
2. Manglish: The test was so hard, I just tembak/tiang only lah!

From the examples above, apart from the underlined word, it is clear that the sentence structure and the usage of ‘lah’ in both sentences are exactly the same. In-depth discussion of the differences between Manglish and Singlish can be found here: https://janneyow.wordpress.com/2014/09/11/singlish-vs-manglish/

B Additional background information

Manglish is developed and used by a multi-ethnic society that is made up of Malay, Chinese, Indians and other ethnic groups. Figure 2 shows one of the theory of how Manglish is developed.

Figure 2: Diagram showing a theory of how Manglish is a result of ‘cross-pollination’ of speakers from different schools or education background. Source: http://mangledmanglish.blogspot.com/2010/10/history-of-manglish.html.

References


