Lecture Notes on Language and Literature (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2026.090104
ISSN 2523-5869 Vol. 9 Num. 1

Improving POS Tagging for Singlish via Data Weighting

Chaojie Lin[1], Xiaoxi Luo[2,*]

[1]Trinity College School, Cambridge, Ontario, Canada
[2]University of Waterloo, Waterloo, Ontario, Canada
**Corresponding author*

Keywords: Singlish, Part-of-Speech Tagging, Low-Resource Languages, Perceptron Tagger, Data Weighting

Abstract: Singlish, or Colloquial Singapore English, is an English-based contact language influenced by multiple substrate languages, including Malay, Tamil, and Southern Chinese varieties. Its mixed grammatical patterns and vocabulary pose significant challenges for standard NLP tools, particularly part-of-speech (POS) tagging. In this study, we investigate whether a simple data-centric strategy—up-weighting Singlish training data while including Standard English UD examples—can improve POS tagging performance without complex architectures. Using an averaged perceptron tagger, we show that the weighted training setup achieves higher accuracy than the Singlish-only baseline, reduces error variance, and successfully captures Chinese-derived grammatical structures. Error analysis indicates that most tagging errors arise from POS polysemy rather than code-mixing, highlighting the effectiveness of data weighting in low-resource settings. Our results suggest that careful data design alone can yield meaningful improvements for processing creole and contact languages.

1. Introduction

Singlish, formally known as Colloquial Singapore English (SCE), is an English-based contact language and creole that developed through long-term multilingual contact in Singapore's sociolinguistic environment[2][5][7] . Singlish differs strongly from Standard English in its grammar and some vocabulary. Its sentence structure and vocabulary are heavily influenced by different languages, such as Malay, Tamil, and Southern Chinese varieties, including Hokkien and Cantonese[5][6]. This mixed structure creates significant challenges for Natural Language Processing (NLP) systems. One key component of NLP is Part-of-Speech (POS) tagging, which automatically assigns grammatical categories to words in a sentence[4]. Most POS taggers are trained on Standard English data and therefore assume English specific grammatical patterns. When these models are applied to Singlish, their accuracy drops significantly[6][7].

Even though Singlish is widely used in daily communication, it has very few annotated linguistic resources. As a result, NLP systems trained only on Standard English often misinterpret Singlish structures and treat them as grammatical errors. Previous research has attempted to address this issue by building Singlish-specific treebanks and using transfer learning methods, such as neural stacking and multi-task learning. These approaches have achieved state-of-the-art POS tagging accuracies of approximately 91.16%[6][7]. While effective, such methods rely on complex architectures and

substantial annotation effort. This raises an important question: to what extent can more straightforward, resource-efficient strategies improve POS tagging performance for Singlish without sacrificing sensitivity to its contact-language grammar?

2. Methodology and Experimental Setup

2.1 POS Tagging Model

In this study, we adopt a perceptron-based POS tagging model implemented in the NLTK toolkit. A Perceptron Tagger is a discriminative sequence labeling model used in Natural Language Processing to automate the assignment of part-of-speech (POS) tags to each word in a text sequence. Unlike generative models, which probabilistically model word–tag sequences, the perceptron is a supervised linear model that learns feature weights to directly map contextual information to the most likely POS tag sequence.

The model uses an averaged perceptron, which stabilizes learning by averaging weight values across all training iterations[3][4]. This technique reduces overfitting and improves robustness, particularly in noisy or low-resource settings such as Singlish.

2.2 Data Representation and Preprocessing

The Singlish data used in this study is drawn from the automatically POS-tagged treebank released by Wang et. al.[6], which was constructed specifically to support computational analysis of Singlish. The dataset is distributed in a CoNLL-style format and provides sentence-level POS annotations suitable for sequence labeling tasks.

The model is trained solely on surface word forms and their corresponding POS tags. No language-specific rules, external lexicons, or handcrafted grammatical constraints are introduced. This design choice ensures that any observed performance differences across experimental conditions can be attributed directly to differences in training data composition rather than feature engineering or rule-based bias.

2.3 Training Procedure and Evaluation Protocol

To account for the stochastic nature of perceptron training, we employ a multi-run evaluation strategy. For each experimental condition, the training data is randomly shuffled at the sentence level prior to training. The model is then trained from scratch and evaluated on a fixed test set. This procedure is repeated ten times using different random seeds.

We report mean tagging accuracy and sample standard deviation across runs. This approach reduces the influence of favorable or unfavorable data ordering and provides a more reliable estimate of model performance. Accuracy is computed as the proportion of correctly tagged tokens over the total number of tokens in the test set.

2.4 Error Analysis

Beyond overall accuracy, we conduct a detailed error analysis to better understand the sources of tagging errors. Model predictions are compared against gold annotations to identify frequent POS tag confusions. In addition, we perform word-level error analysis by tracking the frequency with which individual lexical items are mistagged. This is particularly relevant for Singlish, where many high-frequency words are highly polyfunctional and exhibit different syntactic roles depending on context.

3. Results

The weighted model based on Table 1 achieved 90.03% accuracy, slightly lower than Wang et al.'s reported results.

Table 1. POS tagging accuracy across different training configurations.

Training Configuration	Mean Accuracy	Std. Dev.
Singlish Only (Baseline)	87.14%	± 0.33%
Weighted Singlish (×10) + English UD	90.03%	± 0.21%

4. Error Analysis: Investigating the Sources of Divergence

Table 2. Words with Most Tagging Errors (Weighted Training Set)

Word	Error	Total	Error	Multiple POS Tags?
that	4	12	33.33%	Yes
not	4	26	15.38%	Yes
kay	3	3	100.00%	Yes
of	3	9	33.33%	No
is	3	21	14.29%	No
to	3	44	6.82%	Yes
one	3	10	30.00%	Yes
out	3	4	75.00%	Yes
those	3	4	75.00%	Yes
diam	2	2	100.00%	Yes

Table 3. Words with Most Tagging Errors (Original Training Set)

Word	Error	Total	Error	Multiple POS Tags?
to	9	44	20.45%	Yes
up	5	11	45.45%	Yes
not	4	26	15.38%	Yes
is	4	21	19.05%	No
kay	3	3	100.00%	Yes
are	3	9	33.33%	Yes
better	3	6	50.00%	Yes
%	3	3	100.00%	No
one	3	10	30.00%	Yes
those	3	4	75.00%	Yes

We examined two primary sources of tagging errors on table 2 and table 3:

Claim 1: POS Polysemy. Errors caused by words with multiple valid grammatical categories (e.g., that, one).

Claim 2: Code-Mixing. Errors caused by non-English loanwords or colloquial terms (e.g., diam).

The error distributions in both the weighted and original training sets show a clear and consistent pattern: most tagging errors are concentrated on polysemous English function words, rather than on Singlishor Chinese-derived items. In the weighted training set, the majority of the top error-prone words (e.g., that, not, to, one, out, and those) have multiple possible POS tags, accounting for approximately 80% of the listed high-error items. These words frequently alternate between grammatical categories such as determiners, pronouns, particles, and subordinating conjunctions, making them inherently difficult to disambiguate based on local context. In contrast, high-frequency but less ambiguous words (e.g., is and of) exhibit noticeably lower error rates, indicating that

frequency alone does not explain the observed errors. Code-mixed or Singlish-specific items contribute minimally to the overall error volume: only diam appears among the top errors, and its ambiguity stems from its ability to function as both a verb and an adjective. The recurrence of the same ambiguous function words across both datasets further suggests that the error pattern is systematic rather than data-specific. Taken together, these findings support the conclusion that POS polysemy is the dominant source of tagging errors, while code-mixing plays a comparatively minor role.

5. Experiment 2: Performance on Chinese-Derived Structures

5.1 Criteria for the Chinese-Singlish Dataset

Sentences were selected based on well-documented grammatical features of Singlish. These include sentence-final pragmatic particles, which encode speaker stance or interpersonal meaning[2][5]; the completive marker liao, signaling a change of state or completed action[1]; and the aspectual use of already as a perfective marker rather than a temporal adverb[1]. Additional diagnostics include copula deletion, where the verb to be is omitted before adjectival predicates, and zero-subject constructions, in which subjects are dropped when recoverable from context[5][7].

Table 4. POS tagging accuracy of Chinese-derived tests across different training strategies

Training Configuration	Mean Accuracy	Std. Dev.
Singlish Only (Baseline)	85.58%	± 0.56%
Weighted Singlish (×10) + English UD	88.38%	± 0.18%

5.2 Top Tagging Errors in the Chinese-Singlish Subset

Table 5. Top POS tagging errors for original training set

Word	Error Frequency	Total Frequency
not	8	29
vezel	4	4
better	4	6
that	4	8
fast	3	3
to	3	29
all	3	9
had	2	2
rebus	2	2
honda	2	2

Table 6. Top POS tagging errors for weighted training set

Word	Error Frequency	Total Frequency
not	6	29
better	4	6
vezel	3	4
married	3	3
fast	3	4
to	3	29
that	3	8
had	2	2
dream	2	2
honda	2	2

5.3 Findings

The improvement in accuracy achieved by the weighted dataset table 4 (88.38% vs. 85.58%) further indicates that incorporating English UD data helps the tagger generalize more efectively. Importantly, this gain does not come at the cost of misrepresenting Chinese-influenced Singlish structures. Instead, the model benefits from broader exposure to English function words while still preserving sensitivity to non-standard grammatical patterns.

As can be seen from Table 5 and Table 6, the errors remain concentrated on English function words such as not, to, and that, as well as low-frequency brand names like vezel and honda. In contrast, Chinese-origin particles that are central to Singlish grammar, including lah and liao, do not appear among the most frequent error cases. This suggests that the model is generally able to handle Sinitic-derived discourse particles and aspect markers, even in sentences with strong substrate influence.

Overall, these results support the view that data weighting can strengthen robustness on mixed-language constructions without suppressing substrate-specific features.

6. Discussion

As shown in our results, incorporating English UD data provides broader lexical coverage and more stable representations for high-frequency English function words, which are a major source of tagging errors in Singlish. At the same time, up-weighting the Singlish data by a factor of 10 ensures that Singlish-specific grammatical patterns, including Chinese-derived structures and discourse particles, retain sufficient influence during training.

The efectiveness of this balance is reflected in both overall accuracy and targeted evaluations. The weighted model consistently outperforms the Singlish-only baseline, achieving higher accuracy with lower variance across runs. Importantly, this improvement is also observed in Experiment 2, where the weighted model shows better performance on sentences with strong Chinese substrate influence. Error patterns indicate that Chinese-origin particles such as lah and liao are rarely mis-tagged, suggesting that the inclusion of English data does not override or distort these non-standard structures.

Together, these findings suggest that English data primarily contributes stability and generalization, while data weighting allows Singlish-specific distributions to remain dominant. Rather than competing with each other, Standard English and Singlish data play complementary roles: English supports reliable tagging of common functional elements, and Singlish preserves contact-induced grammatical features. This highlights data weighting as a simple yet efective strategy for modeling low-resource contact varieties.

7. Conclusion

This study examined whether simple data-based methods can improve POS tagging for Singlish, a low-resource language variety with strong influence from other languages. Using an averaged perceptron tagger, we showed that giving more weight to Singlish data while adding standard English UD data leads to clear and stable improvements in accuracy. This was done without using language-specific rules or complex model designs.

Across all experiments, the weighted training setup performed better than the Singlishonly baseline. Although the final accuracy is slightly lower than the result reported by Wang et al., the performance remains competitive. The low variation across repeated runs shows that the improvements are reliable and not caused by randomness in training. Error analysis shows that most tagging mistakes come from POS polysemy, meaning that some words can belong to more than one grammatical category. Common words such as that, one, to, and out appear in many diferent sentence

roles, which makes them hard to tag correctly. In contrast, errors caused by code-mixing are much less common. Words borrowed from Chinese dialects, such as lah, liao, and diam, rarely appear among the most frequent errors.

This pattern is also seen in Experiment 2, which focuses on sentences with strong Chinese influence. The weighted model achieved higher accuracy than the Singlish-only model on this subset. Importantly, most errors still involved English function words or rare proper names, not Chinese-derived particles. This suggests that the model can successfully learn the grammar of Chinese-influenced Singlish structures.

Overall, the results show that standard English data can help stabilize learning, while data weighting allows Singlish-specific patterns to remain important. For low-resource language varieties, this approach ofers a simple and efficient alternative to complex neural models, showing that careful data design alone can lead to meaningful improvements.

Acknowledgement

Future research can explore the optimal weighting ratio (e.g., 5x vs. 20x) and apply this methodology to other English-based languages. This will help assess the generalizability of data weighting as a strategy for modeling substrate efects in diverse varieties of Englishbased languages.

References

[1] Bao, Z. (2005). The aspectual system of Singapore English and the systemic substratist explanation. Journal of Linguistics, 41(2), 237–267. https://doi.org/10.1017/S0022226705003269

[2] Gupta, A. F. (1992). The pragmatic particles of Singapore Colloquial English. Journal of Pragmatics, 18(1), 31–57. https://doi.org/10.1016/0378-2166(92)90106-L

[3]Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1373–1378). Association for Computational Linguistics.

[4] Jurafsky, D., & Martin, J. H. (2025). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models (3rd ed.). Online manuscript. Stanford University.

[5] Lim, L. (2004). Singapore English. John Benjamins Publishing Company. https://doi.org/10.1075/veaw.g33

[6] Wang, H., Yang, J., & Zhang, Y. (2019). From genesis to creole language. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(1), 1–29. https://doi.org/10.1145/3321128

[7] Wang, H., Zhang, Y., Chan, G. L., Yang, J., & Chieu, H. L. (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) (pp. 1732–1744). Association for Computational Linguistics.