POS Tagging Model for Malay Tweets Using New POS Tagset and BiLTSM-CRF Approach

Sabrina Tiun [1], Siti Noor Allia Noor Ariffin [1] and Yee Dhong Chew [2]

[1] *ASLAN lab, Center of Artificial Intelligence, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 45600 Bangi, Selangor*

[2] *NETS Solutions PTE LTD, 19th Floor – Wisma Lee Rubber, 50100 Kuala Lumpur*

Abstract

This paper proposes a Malay Part-of-Speech (POS) tagger using a new set of POS tags and a deep learning classifier. A new set of POS tags was proposed, and the POS classifier was built using a deep learning model, a bidirectional long short-term memory network with a conditional random field (BiLSTM-CRF). We expanded the POS tagset by considering the informal Malay terms frequently used in Malay tweet text. Additionally, for the Bi-LTSM-CRF model, we used a combined embedding of Malay Wiki Word2Vec and a Malay Twitter corpus annotated with POS Word2Vec. The BiLSTM-CRF Malay POS tagger was then compared to traditional classifier models. Based on evaluation results, the BiLSTM-CRF model performed the best. Thus, we concluded that deep learning techniques combined with appropriate embeddings are capable of performing fine-grained POS tagging on Malay tweet text. Additionally, using a customized POS tagset for specific types of text results in increased POS label coverage.

Keywords

Part-of-speech tagging, Malay POS tagger, Malay Tweets, BiLSTM-CRF

1. Introduction

The Part-of-Speech (POS) classification method organizes words into groups based on how they are used and function in sentences. A POS tagger, on the other hand, is a software component that reads a text in many languages and assigns appropriate words to each word (or token) in the text. Malay has four primary POS tags: Nouns, verbs, adjectives, and word tasks [1]. On the other hand, these leading POS tags are better suited for tagging formal Malay language than informal Malay material, such as Malay Twitter data. Informal Malay contains a varied range of informal terminology. The varied range includes accent (or dialect) words, slang, titles (e.g., *hang, mek*), noises (words intended to describe sounds like laughter, cat purring, and knocking), and mixed languages (commonly a mixed of a Malay language with the English language).

Additionally, Malaysians, particularly adolescents, are incredibly inventive when it comes to writing and coining new words [5]. As a result, Malay social media text, such as Malay tweets, is dense with colloquial Malay and peppered with mixed-language phrases and derogatory terms. As a result, POS tagging a Malay tweet is extremely complicated, time-consuming, and energy-intensive.

Obtaining the POS tag to label these informal terms is one of the challenges. Some researchers either ignore these terms (by failing to label them) or repurpose the existing POS tagset. Both approaches reduce the accuracy of POS tagging on tweet text and cause POS tags to be mislabeled. To address this issue, we propose a new set of POS tags customized for Malay social media text, the tweet. On the other

The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), June 7-8, 2022, Koper, Slovenia

EMAIL: sabrinatiun@ukm.edu.my (Sabrina Tiun); p102054@siswa.ukm.edu.my (Siti Noor Allia Noor Ariffin), A170527@siswa.ukm.edu.my (Yee Dhong Chew)

hand, the POS tagging problem is highly dependent on the preceding and following sequences. Applying a neural network algorithm, such as a bidirectional long short-term memory network (BiLSTM) model to POS tagging will be extremely beneficial. In other words, the architecture of the BiLSTM model, which considers both previous and current sequences, increases the likelihood of constructing a high-accuracy prediction of the POS tag on the Malay tweet. Thus, to develop a high-performance POS tagger for Malay tweets, we propose a new POS tagset and a BiLSTM with Conditional Random Field (CRF) or BiLSTM-CRF model.

2. Research Method

Our work in this paper consists of two parts: (1) to propose a finer POS tagset that covers nearly all types of words in the tweet text. (2) to build a POS tag classifier model based on the BiLSTM-CRF model. The following will describe in detail of our work:

2.1. Malay POS Tagset for Tweet Text

Malay has four primary POS tags, but these are insufficient for categorizing words in the Malay Twitter corpus. As a result, we chose to create new Malay POS tags through the use of Malay formulas and the grammar of Safiah et al. [1]. Additionally, we modified the newly created POS tags to conform to the Malay Twitter data criteria by comparing them to the word classes identified in Safiah et al. [1] and Othman and Karim [2], as well as some prior findings on social media texts [3][4]. Furthermore, we discovered that several of the newly developed POS tags by Le et al. [3] are suitable for categorizing words in the Malay tweet. For instance, consider the FOR and NEG tags. The FOR tag is used to categorize foreign language words found in the study corpus. On the other hand, the NEG tag identifies words with negative connotations, such as those that refer to swearing.

As a result, we chose to use Le et al. [3] 's two newly developed POS tags, namely FOR and NEG tags. The FOR and NEG tags can be used for Malay tweets since Malaysians enjoy combining words from multiple languages in tweets and using negative words to express emotions or disapproval of situations. In addition, removing a foreign language from the corpus changes the author's intended meaning as well as the writing structure of the tweets. When auto annotations and annotators attempt to tag the POS tags due to this change, an error will occur. In that case, we will not exclude foreign language terms, slang terms, or informal Malay expressions that follow this principle—instead, a unique POS tag designed for this type of word tagging will be used. Thus, we combine lists of POS tagset from the following sources: Safiah et al. [1], Othman and Karim [2], Le et al. [3] and Ariffin and Tiun [4] and our 18 new POS tagset. In other words, the new set of POS tags contains 45 POS tags, and the detail of the POS tagset can be seen in Appendix A.

2.2. BiLTSM-CRF Model for Malay POS tagging

For the POS tag classifier model, we constructed our Malay POS tagging model based on BiLSTM-CRF. The model is based on these three phases: (1) data preparation, (2) building/training the BiLSTM-CRF Malay POS model, and (3) model evaluation. 30% (538) of the total data was set as test data in the data preparation. When training the biLSTM-CRF Malay POS model, three kinds of layers were built: The embedding layer, the BiLTSM layer, and the CRF layer (see the Figure 1):

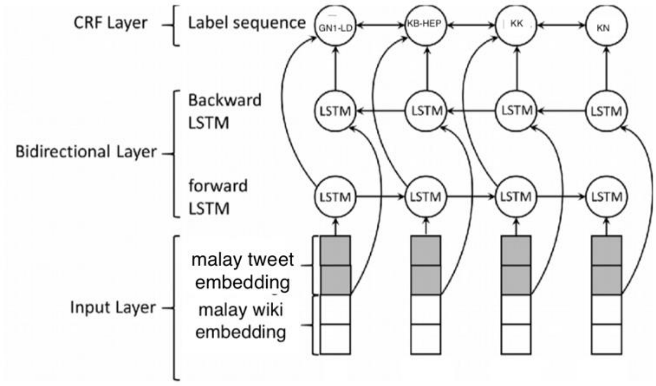

Figure 1: The BiLSTM-CRF Malay POS model (adapted from Zhang et al. [12])

At the *embedding layer*, a combination of embeddings was proposed; the Word2Vec of Malay Wikipedia [11] and the Word2vec of an annotated POS Malay tweets corpus. Both of the embeddings were trained with the parameter setting of 300 dimensions. As for the *BiLTSM layer*, the BiLTSM acts as a word feature extraction. Bi-LSTM takes contextual information and identifies class tagging potentials for each input. Finally, the *CRF layer* is used to classify the POS of a word. The CRF score function is used to find the POS sequence with the highest score and calculate the probability distribution of all POS sequences. The CRF layer benefits from knowing what word should be followed by what word. Such features are difficult to demonstrate in the neural network layer, especially on small data sets [6]. However, with the help of the CRF layer, even a model trained with a small dataset can be used to classify things well.

Using the prepared dataset, we evaluated our BiLSTM-CRF model. In evaluating our model, a similar evaluation of Jason [12] that used training loss and validation loss were used to ensure our BiLSTM-CRF model was not underfit or overfit. Based on Figure 2 below, the small gap in learning curves between training loss and validation loss indicates the model was neither underfitted nor overfitted.

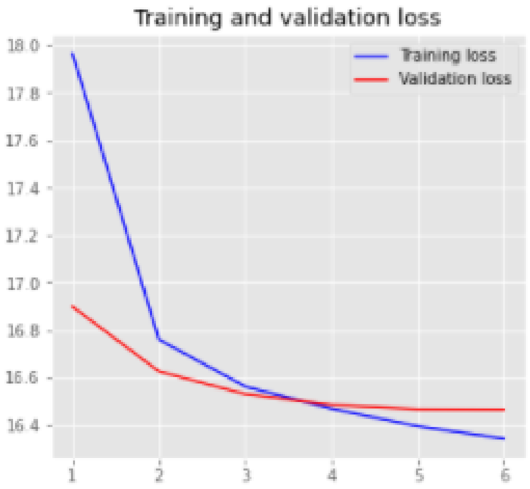

Figure 2: Learning curves of the BiLSTM-CRF Malay POS model

Afterwards, we ran several trainings using a set of hyperparameters. The most optimized hyperparameters for our BiLTSM-CRF Malay POS model were; learning algorithm Adam, dropout point 0, batch size 6, and 30 epochs. With such setting, our BiLSTM-CRF model managed to get 94% of Precision, Recall and F1-Score.

3. Result and discussion

To evaluate our POS model, we compared it against four well-known traditional classifier models: Support Vector Machine (SVM)[7], Nave Bayes (NB)[9], Decision Tree (DT)[8] and K-nearest neighbour (KNN)[10]. All of the traditional models were trained based on word position features. To be precise, ten types of features were extracted: the preceding and following words; the prefix of each word (limited to the first three characters of the word); the suffix of each word (limited to the last three characters of the word); the length of the word; and the presence of a digit in the word. To train and evaluate traditional classifier models, the same dataset of Malay tweets (used in training and evaluate the BiLSTM-CRF mode) was used and divided into 70% (1,253 data) as trained data and 30% (538 data) as test data with 10-fold cross validation. To compare our BiLSTM-CRF model against the traditional classifier, evaluation metrics of precision, recall, and F1-score were used. The findings based on the metrics are shown in Table 1.

Table 1
Result of BiLSTM-CRF classifier against traditional classifiers on tagging Malay POS

Classifier	Precision	Recall	F1-score
SVM	0.93	0.91	0.92
NB	0.74	0.57	0.60
DT	0.89	0.87	0.89
KNN	0.85	0.88	0.85
BiLTSM-CRF	0.94	0.94	0.94

Though traditional and deep learning classifiers have distinct architectures (see Table 1), using the same dataset gives us a general view of the Malay POS BiLSTSM-CRF model's performance. Given SVM classifier has an F1-Score of 92% compared to the BiLTSM- CRF with an F1-Score of 94%, the 2% difference is small. The results in Table 1 can probably be improved by employing SVM with more significant features, or the BiLTSM-CRF will perform better with a larger dataset. However, because BiLTSM-CRF has been shown to perform better with a large dataset, strengthening the BiLTSM-CRF model in the future will be a preferable option for future study.

4. Conclusion

In conclusion, by creating new POS tags tailored to the words in Malay tweets, we ensure fewer unlabeled words with POS. The training of the BiLTSM-CRF model based on a combination of specific embeddings aids in increasing the performance in tagging POS in the tweet text. In other words, our proposed BiLSTM-CRF Malay POS tagging model and the new POS tagset are suitable for tagging POS in the Malay tweet text.

5. Acknowledgements

This project is supported by the Malaysia Ministry of Higher Education under research code: FRGS/1/2020/ICT02/UKM/02/1

6. References

[1] K. N. Safiah, F. M. Onn, H. H. Musa, A. H. Mahmood, Tatabahasa Dewan Edisi Ketiga, Dewan Bahasa dan Pustaka, Kuala Lumpur, 2010.

[2] A. Othman, N. S. Karim, Kamus komprehensif Bahasa Melayu, Penerbit Fajar Bakti, Kuala Lumpur, 2005.

[3] T. A. Le, D. Moeljadi, Y. Miura, T. Ohkuma, Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets, in: Proceedings of the 12th Workshop on Asian Language Resources, Osaka, Japan, 2016, pp. 123-131.

[4] S. N. A. N. Ariffin, S. Tiun, Part-of-Speech Tagger for Malay Social Media Texts, GEMA Online® Journal of Language Studies, 18 (2018) 124 -142.

[5] N. Jamali, Fenomena Penggunaan Bahasa Slanga dalam Kalangan Remaja Felda di Gugusan Felda Taib Andak: Suatu Tinjauan Sosiolinguistik, Jurnal Wacana Sarjana, 2 (2018) 1-1.

[6] L. Marz, D., Trautmann, B. Roth, Domain Adaptation for Part-of-Speech Tagging of Noisy User-Generated Text, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2019.

[7] T. Nakagawa, T. Kudoh, Y. Matsumoto, 2001. Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines, in: Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, 2001, pp. 325--331.

[8] L. Màrquez, H. Rodríguez, Part-of-speech tagging using decision trees, in: Proceedings of the European Conference on Machine Learning, 1998, pp. 25-36.

[9] R. Creţulescu, A. David, D. Morariu, L. Vinţan, Part of speech tagging with Naïve Bayes methods, in: Proceedings of the 18th International Conference on System Theory, Control and Computing (ICSTCC), 2014, pp. 446-451.

[10] S. Gaber, N., Nazri A. M. Z., Omar, N., Abdullah, S., Part-of-Speech (POS) Tagger for Malay Language using Naïve Bayes and K-Nearest Neighbor Model. International Journal of Psychosocial Rehabilitation, 24 (2020) 5468-5476.

[11] Malaya toolkit, Natural Language Toolkit library for Bahasa Melayu, 2020. URL: https://malaya.readthedocs.io/en/4.0/load-wordvector.html

[12] Y. Zhang, X. Wang, Z. Hou, J. Li, Clinical named entity recognition from Chinese electronic health records via machine learning methods, JMIR Medical Informatics, 6 (2018) e50.

Appendix A

The proposed new POS tagset for Malay informal text which are the combination of the POS tagset from Safiah et al. [1], Othman dan Karim [2], Le et al. [3] dan Ariffin dan Tiun [4] and a new proposed POS. The bold POS depicts the new proposed POS.

POS tag	*Description*
KN	Noun (e.g kereta 'car')
KN-LD	Noun-dialect (e.g gerek 'bike')
KN-KEP	Noun-abbreviation (e.g keta 'car')
GT	Pronoun preposition (e.g sini 'here')
GT-KEP	Pronoun preposition-abbreviation (e.g ni 'here')
GDT-KTY	Pronoun-question (e.g siapa 'who')
GN1	1st person personal pronoun (e.g saya' me')
GN1-LD	1st person personal pronoun-dialect (e.g cheq 'me')
GN2	2nd person personal pronoun (e.g awak 'you')
GN2-LD	2nd person personal pronoun-dialect (e.g hang 'you')
GN3	3rd person personal pronoun (e.g mereka 'they')
GN3-LD	3rd person personal pronoun-dialect (e.g depa 'they')
KK	Verb (e.g lari 'run')
KA	Adjective (e.g dekat'near')
KA-KEP	Adjective abbreviation (e.g kat 'near)
KH	Conjuction (e.g dan 'and')
KH-KEP	Conjuction abbreviation (e.g tapi 'but')
KSR	Interjection (e.g wah)
KTY	Question (e.g bila 'when')
KPE	Command (e.g sila 'please')
KB	Auxiliary verb (e.g akan 'will')
KB-KEP	Auxiliary verb abbreviation (e.g dah)
KP	*Kata penguat*/adverb (e.g sangat ‘very’, paling ‘very’)
KPN	*Kata Penegas*/adverb (e.g juga ‘too’)
KNF	Deny (e.g tidak 'no')
KNF-KEP	Deny abbreviation (e.g tak 'no')
KPM	Narrator (e.g ialah 'is')
KS	Noun preposition (e.g di 'at')
KPB	Justified word (e.g ya 'yes')
KBIL	Cardinal (e.g ribu 'thousand')
KAR	Preposition (e.g atas 'above')
KAD	Adverb (e.g sekarang ‘now’)
KAD-KEP	Adverb abbreviation (e.g dulu)
FOR	Foreign word (e.g I 'I'')
FOR-KEP	Foreign word abbreviation (e.g iols 'we')
FOR-NEG	Foreign word - Negative (e.g b*shit)
NEG	Negative (e.g bodoh 'stupid')
KD	Preposition (e.g di 'at')
MW	Currency (e.g MYR)
LD	Dialect (e.g
SL	Slang (e.g pastu 'after that')
KEP	Abbreviation (e.g pi as pergi 'go')
GL	Title (e,g hang)
BY	Sound (non-speech)
AWL	Prefix (e.g anti-)