Benchmarking Mi-POS: Malay Part-of-Speech Tagger

Benjamin Chu Min Xian, Mohamed Lubani, Liew Kwei Ping, Khalil Bouzekri, Rohana Mahmud, and Dickson Lukose

***Abstract*—A part-of-speech tagger as signs the correct grammatical category to each word in a given text based on the context surrounding the word. This paper presents Mi-POS, a Malay language Part-of-Speech tagger that is developed using a probabilistic approach with information about the context. The results of benchmarking Mi-POS against several similar systems are also presented in this paper and the lessons learnt from it are highlighted. The dataset used for evaluation consists of manually annotated texts. The authors used the accuracy and time to measure the results of this evaluation. The final results show that Mi-POS outperforms other Malay Part-of-Speech taggers in terms of accuracy with an accuracy of 95.16% obtained by tagging new words from the same training corpus type and 81.12% for words from different corpora types.**

***Index Terms*—Benchmarking, Malay language, natural language processing, part-of-speech tagging.**

I. Introduction

Part-of-speech (POS) tagging is an important process that is used to build many Natural Language Processing (NLP) applications. The POS tagger assigns a unique grammatical class to each word in a context (e.g., a sentence). However, natural language words can have different POS tags based on their contexts. This ambiguity makes the POS tagging a non-trivial process since context interpretation is essential to find the correct tag for a given word. To automate this process, machine learning techniques including statistical and probabilistic methods have been used to build powerful POS taggers.

Training the machine learning models necessitates a manually-built POS-tagged corpus to be able to predict the correct tags for new words. Such corpus may be available for the major languages. However, due to the lack of linguistic resources for Malay language, this corpus needs to be constructed manually to be used to train the POS models.

In this paper, a Malay POS tagger called Mi-POS is developed and compared with other existing Malay POS taggers. A manually-built corpus is constructed to train the models. Another two manually-built corpora are used to test the models.

To this end, this paper is structured as follows: Section II describes the related work on existing POS taggers; Section III highlights the proposed model of our Mi-POS; Section IV shows all the results of the experiment; Section V discusses the results and performance of Mi-POS compared to other systems. Finally, Section VI concludes this paper with a discussion on the overall outcome achieved and future research directions.

II. Related Work

POS tagging is widely adopted for languages such as English, German, Spanish and Arabic [1]-[4]. It plays a significant role in text analysis as it is an initial step to identify the grammar information in the text. Among the existing POS taggers are TnT Tagger [2] and Brill Tagger [5]. All of them are adopting machine learning methods and achieve accuracies of 96.7%, 97.24% and 95% respectively. The rich availability of linguistic resources is the main factor which contributes to the development of these taggers<for the European languages. However, in contrast, there is less research on POS for Malay language due to its limited resources.

One Malay Tagger is developed by Mohamed [6] which applies trigram Hidden Markov Model (HMM) method to identify words' tags in Malay sentences. Context information other than the surrounding tags, namely the prefix and the suffix, has been used to predict the correct POS tags. His study measures the effect of using these features individually as well as using a combination of both the prefix and the suffix of each word in the final model's predictions. The model is tested using a corpus of 18,135 tokens tagged with a set of 21 tags similar to the set of tags used by Dewan Bahasadan Pustaka (DBP) [7]. This corpus is tagged automatically by mapping each word to a list of possible tags from a dictionary, and then the ambiguity is solved manually. The results show that the best predictions are made with accuracy 67.9% using only prefixes information with a fixed prefix length equals to three letters. Similar results with accuracy 66.7% are achieved using a combination of the first and the last three letters of each word. When using suffixes information only, the best accuracy achieved is 60% with suffix length of five letters. These findings show that HMMs are suitable models to be used to predict any Malay word's POS tag.

On the other hand, Rayner Alfred *et al*. proposed a rule-based method for identifying Malay POS tags called RPOS [8]. It applies affixing and word relation rules to determine the right word category. Malay words can be formed with prefixes, suffixes, circum fixes and/or infixes. In

Manuscript received December 12, 2015; revised February 29, 2016.

Dickson Lukose, Khalil Bouzekri and Benjamin Chu Min Xian are with the Artificial Intelligence Lab at MIMOS Berhad, Kuala Lumpur, 57000 Malaysia (e-mail: dickson.lukose@mimos.my, khalil.ben@mimos.my, mx.chu@mimos.my).

Mohamed Lubani and Liew Kwei Ping are with the University of Malaya, Faculty of Computer Science and Information Technology, Kuala Lumpur, 50603 Malaysia (e-mail: mohamed.lubani@siswa.um.edu.my, liewkweiping@siswa.um.edu.my).

Rohana Mahmud is with the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603 Malaysia (e-mail: rohanamahmud@um.edu.my).

their paper the authors consider infixes less important and not effective for the task of POS tagging. Different affixes can be categorized in different word categories. For example, the verb type can be identified by the prefix *mem-* while the noun type involves prefixes such as*pen-*. When there is more than one possible tag for the word, word relation rules are applied to identify the most suitable POS tag based on the context. If the word is not found in the POS tag dictionary, affixing rules are applied to determine possible tags for the word and then the word relation rules are applied to solve the ambiguity (if any). The POS dictionary is manually built from Thesaurus Bahasa Melayu [9] and used to assign all possible tags to each word in a Malay sentence. The results of this rule-based method show that it has higher performance than the statistical POS tagger with an accuracy of 89% for Malay news articles and 86% for Malay biomedical articles. This shows that it is able to predict unknown words' POS tags at a reasonable accuracy. However, this tagger fails to tag words that are borrowed from English and also words that have no affixation especially proper nouns. Richer relation rules are needed to improve the tagging results of RPOS tagger.

A statistical unsupervised machine learning method for POS tagging was introduced by Norshuhani Zamin *et al.* as "Lazy Man's Way" method [10]. It is called "Lazy Man" because it does not require laborious effort to annotate the Malay dataset for training purposes like taggers in [6]-[8], and [11]. It annotates Malay sentences with POS tags using a Malay-English lexicon. First, Malay words are translated into English using Google Translate. The English words are then tagged using Brill' stagger [5]. After that, the results are mapped with the help of the Malay-English lexicon using N-gram and Dice Coefficient approaches for similarity measurement purposes. The system has a precision of 86.87% and a recall of 72.56%. This system can be used when there is no or limited language resources. However, since grammatical structure is different between Malay and English, the system may produce inaccurate results if it relies only on the Malay-English lexicon. Therefore, large set of lexical and disambiguation rules are needed to improve the system.

A relational lexical database, MALEX (MALayLEXicon) with purpose of providing linguistic information for Malay text analysis, has been developed by Gerry Knowles and Zuraidah Mohd Don [11]. They emphasize syntactic drift approach and data-driven approach to identify the Malay grammar class [11], [12] where the tag is recognized through examining the syntactic structure. For example, the word "keras" (hard) has different tags. In the sentence "Buahini *keras*" (The fruit is hard), the word "keras" is an adjective whereas in the sentence "Buahitusudah *keras* sekarang" (The fruit has become hard now), which contains a time marker "sudah", "keras" is considered as a verb. It can also be considered as an adverb as in the sentence "Kami berker ja*keras*" (We work hard) [11]. It integrates normalizing, stemming, tagging, parsing and pronouncing to identify the Malay tag. The researchers have further performed a corpus-based approach to analyze the Malay grammatical class. Around 120,000 words from four DBP novels have been manually tagged and used for this supervised system. The annotated corpus is used to predict unseen words in new dataset [10]. However, there is a lack of technical discussion on the learning approach and the performance of the system remains unclear.

Another POS tagger proposed in [13] shows possible methods for building a POS tagger for Bahasa Indonesia. Two machine learning methods namely Conditional Random Fields (CRF) and Maximum Entropy (MaxEnt) are compared using two different corpora. A CRF model consists of several feature functions weighted with values learned from the training corpus. This makes CRFs more general than HMMs and not restricted by the independence assumptions found in HMMs. In HMMs, each word depends on the current tag and each tag depends only on the previous tags. Also CRFs are not limited to constant probabilities since each feature function is weighted with unlimited weight value. Therefore, using CRFs to build a POS tagger is more preferable for the previous reasons. The second method used in [13] is MaxEnt to build a flexible POS tagger that relaxes probability assumptions and maximizes the use of context information to find the model with the highest entropy value of the probability distribution of the training data. For evaluation, [13] uses two different corpora tagged manually using two different tag sets with 37 tags and 25 tags. The first corpus is a manually-built corpus consists of 14,165 tokens, and the second is part of the Pan Localization project and considered as the equivalent of Penn Tree Bank corpus for Bahasa Indonesia with approximately 500,000 tokens. Results show that MaxEnt model has the highest accuracy for both tag sets in the two corpora with a better average accuracy when using the 25 tag sets which measured at 85.02% for the first corpus and 97.57% for the second corpus. The best accuracy for the CRF model is 91.15% using the 25 tags on the second corpus. Table I shows a summary of the comparison between the techniques mentioned

TABLE I: COMPARISON BETWEEN MALAY POS TAGGERS

Method	Strengths	Weaknesses
Trigram HMM [6]	• Solid Theoretical basis. • Utilizes context information.	• Expensive computations. • Large training data.
RPOS [8]	• Able to predict unknown word's POS tag without training data. • High performance for main POS categories.	• Requires sufficient number of rules. • Laborious effort. • Low accuracy for non-rule rich POS categories.
MALEX [12]	• Tagging based on syntax. • Comprehensive data.	• Laborious effort.
Lazy Man's Way [10]	• Fast as no laborious effort is required. • Can be used when there is no or limited language resources.	• Different language structures lead to incorrect mapping. • Low accuracy.
Probabilistic Models (CRF & MaxEnt) [13]	• General than HMMs. • Relaxed probability assumptions.	• Complex structures. • Require additional feature functions.

The ambiguity issue is the most challenging part in text analysis for POS tagging. The same word may have different meanings in different contexts. Most of the Malay POS

taggers identify the tag category based only on the word itself and not based on the context. This may result in improper text analysis and thus inaccurate POS tags. Therefore, a context-based POS tagger which identifies word categories based on the meaning of the sentence is necessary.

III. Part-of-Speech Model

Mi-POS is inspired by the probabilistic methods to perform the task of POS tagging. The MaxEnt method is used to develop Mi-POS where the joint probability of a specific class c assigned to a token t based on their occurrences in a training corpus is used. The estimated value $P(c,t)$ is expected to maximize the entropy function defined in Equation (1).

$$H(P) = -\sum_{(c,t)\in C\times T} P(c,t)\, log\big(P(c,t)\big) \qquad (1)$$

where C is the set of all classes, T is the set of all unique tokens.

OpenNLP [14] is an open source NLP code library with pre-trained models to perform different NLP tasks such as tokenization, POS tagging, Named-Entity recognition (NER), chunking, parsing and coreference resolution. Although there are pre-trained POS models available in OpenNLP for selected languages, there is no model available for Malay language. For the development of Mi-POS as a POS tagger for Malay language, we use OpenNLP and our manually built training corpus with Bernama news archive [15]. The following section describes the process of transforming an unlabeled corpus to a POS annotated corpus which will be used as the input to OpenNLP to generate the Malay POS model.

A. Dataset

The annotated training data is a collection of tokenized sentences where each token has an assigned POS tag. This manually-tagged corpus contains 152 articles with a total of 64,534 tokens from Bernama news archive. It was manually tagged by a Malay native speaker who assigned a single POS tag to each word. Then the tagged corpus was verified by two other Malay native speakers to correct mistakes and solve ambiguity if any.

There is a total of 13 non-symbols POS types used to tag the training corpus by the Malay native speakers, which can be referred in Table V.

The datasets that are used in existing Malay Taggers [6], [8], [11] to perform the evaluation are limited with 18,135 tokens, 8,700 tokens and 120,000 tokens respectively and these datasets are not publicly accessible. Therefore, to evaluate the performance of the proposed Malay POS tagger Mi-POS, another two manually-tagged Malay datasets are used. Three Malay native speakers had been invited to build the manually-tagged datasets for testing the model. The two Malay datasets consist of 40 different articles with 20 articles in each dataset and an average of 359 tokens per article for the first corpus and 550 tokens per article in the secod corpus. The articles in the first dataset are taken from Bernama news archive whereas the second dataset contains selected articles from the Malay corpus developed in [16] by Su'ad Awab which contains different categories including art, economics, education, health, information technology, low, literature, sport, and science. All categories are included when selecting the articles for the second dataset.

The two testing corpora are built similar to building the training corpus where the first tagging phase is done by one Malay native speaker to assign the correct tag / tags to each word. After that, the results are checked and verified by two other Malay native speakers to assign one tag to each word. The final result is used as our gold standard to evaluate the performance of the proposed Malay POS Tagger.

B. Training

In order to use the Bernamanews archive to train the model, we first converted it to the OpenNLP format by placing each sentence on a separate line. The end of a sentence is marked with "._.". The word tokens are combined with an underscore"_" and followed by the corresponding POS tag, whether the word is a noun (NN), verb (VB), preposition (IN) or adjective (JJ) as shown in Fig. 1.

```
Dia_PRPpergi_VBke_INrumah_NNkawannya_NNpada_INhujung_JJ
minggu_NN ._.
```

Fig.1. Sample annotated sentence.

All of the annotated sentences are stored in a training file called "ms-pos.train". We use the command line shown in Fig. 2 to generate the Malay POS model, with the setting of the parameters for the language to Malay (ms), the model type to "maxent", input data to "ms-pos.train" and the output model file to "ms-pos-maxent.bin".

```
$bin/opennlpPOSTaggerTrainer -encoding UTF-8 -langms -model-type
maxent -data ms-pos.train \
-model ms-pos-maxent.bin
```

Fig. 2. OpenNLP command line.

IV. Experiments and Results

The results of Mi-POS are compared with the statistical Malay POS taggers proposed in [6] using trigram HMM. The best two HMMs used in the paper, i.e., the trigram HMM using prefixes only smoothed by successive abstraction and the unsmoothed trigram HMM using both prefixes and suffixes, are used in our comparative study.

The training accuracy of Mi-POS and the HMMs, which is obtained by calculating correct tags ratio of these systems tested on the same training data, is shown in Table II.

TABLE II: Training Accuracy for Mi-POS and the Two Used HMMs

	Mi-POS	Trigram HMM using prefixes (Linear smoothed)	Trigram HMM using both prefixes and suffixes
Training Accuracy	99.61%	93.4%	94.87%

Table III shows the results of Mi-POS when tested using both testing corpora. An accuracy of 95.16% is obtained for the first corpus with news articles and 81.12% when using the second non-news corpus. The processing time of Mi-POS is also shown in the same table for both testing corpora.

Table IV shows the results of the two used HMMs using the two different testing corpora. The accuracy of a trigram

HMM using information of both the prefix and the suffix is higher than that of a trigram HMM using only prefix information for both the testing corpora. Mi-POS outperforms both HMM models with higher accuracy and significantly better processing time.

TABLE III: TAGGING ACCURACY FOR NEWS AND NON-NEWS ARTICLES FOR MI-POS

	Tokens (A)	Correctly Tagged (B)	Accuracy $(\frac{B}{A} \times 100)\%$	Time (seconds)
News Articles (first corpus)	7170	6823	95.16%	5.17
Non-News Articles (second corpus)	10989	8104	81.12%	7.80

TABLE IV: TAGGING ACCURACY FOR NEWS AND NON-NEWS ARTICLES BY THE TWO HMMS

	Tokens (A)	Correctly Tagged (B)	Accuracy $(\frac{B}{A} \times 100)\%$	Time (seconds)
Trigram HMM using prefixes (Linear smoothed)				
News Articles (first corpus)	7170	5844	81.51%	851
Non-News Articles (second corpus)	10989	7067	64.31%	1,044
Trigram HMM using both prefixes and suffixes				
News Articles (first corpus)	7170	5944	82.9%	116
Non-News Articles (second corpus)	10989	7049	64.14%	136

The results of Mi-POS are also compared with the results of the unsupervised tagger "Lazy Man" proposed in [10]. For this purpose, sentences in each of the two corpora were first translated into English using the online Google Translate service. Then, the translated sentences were passed to the free online version of Brill's tagger [17] to be annotated with POS tags. And since Brill's tagger has its own text tokenization algorithm, we enclose sentences in Malay version and English version with a special marker to make sure that all sentences are aligned and not further split by Brill's tagger. To build the Malay-English lexicon, a special Web crawler was built to extract entries from the freely online available 'Malay-English dictionary' [18]. We collected 18,177 unique Malay entries with their corresponding English translations. To enhance the dictionary lookup process, we first lookup the Malay word in its original form, if not found in the dictionary, the lemma of that word is used. Malay words other than symbols that cannot be found in the dictionary will be tagged with "NA".

The Lazy Man's tagger uses a mapping process between the Malay sentence and the corresponding tagged English sentence. The processing time of this mapping can be ignored compared to the time required to translate the sentence and apply the Brill's tagger. The processing time of the Lazy Man's tagger depends on external online services and, therefore, cannot be determined with full accuracy.

To correctly compare the results of Mi-POS with the Lazy Man's tagger, both POS tag sets need to be mapped to the same tag set. Table V shows the tags generated by Brill's tagger which is used in the Lazy Man's tagger and their Mi-POS corresponding tags.

Table VI shows the results of the Lazy Man's tagger for both the test corpora. It is noticed that the accuracy of this tagger is lower than that of Mi-POS and HMMs.

TABLE V: LAZY MAN'S TAGS AND THEIR MI-POS CORRESPONDING TAGS

Lazy Man's Tags	**Mi-POS corresponding tags**	**Mi-POS tags definitions**
CC	CC	Coordinate conjunctions (ex. dan, atau)
CD, PDT	CD	Cardinals (ex. satu, juta, kedua)
DT	DT	Determiners (ex. ini, itu)
EX, RB, RBR, RBS, RP	RB	Adverbs (ex. sekarang)
FW, NNP, NNPS	NNP	Proper nouns (ex. Africa, KL)
IN, TO	IN	Prepositions (ex. di, ke, dari)
JJ, JJR, JJS	JJ	Adjectives (ex. kecil, besar)
LS	*Ignored*	-
Ignored	NEG	Negations (ex. bukan, tidak)
MD, VB, VBD, VBG, VBN, VBP, VBZ	VB	Verbs (ex. berlari, membaca,akan)
NN, NNS	NN	Nouns (ex. kitab, orang)
POS, PRP, PRP$	PRP	Pronouns (ex. saya, mereka)
UH	UH	Interjections (ex. aduh, oh)
WDT, WP, WP$, WRB	WH	WH(ex. apa, siapa)

TABLE VI: TAGGING ACCURACY FOR NEWS AND NON-NEWS ARTICLES FOR THE LAZY MAN'S TAGGER

	Tokens (A)	Correctly Tagged (B)	Accuracy $(\frac{B}{A} \times 100)\%$	Time (seconds)
News Articles (first corpus)	7170	4139	57.73%	≈ 6.7
Non-News Articles (second corpus)	10989	5747	52.3%	

In our efforts to benchmark the results of Mi-POS tagger, we also implemented the rule-based POS tagger RPOS proposed in [8]. There are two sets of rules used to implement the RPOS tagger: the word relation rules and the affixing rules. There are 14 different POS tags that could result after applying the word relation rules. From these rules we choose only the POS tags that exist in the tag set of Mi-POS. These POS tags are: noun, verb, adjective, adverb, preposition, conjunction and cardinal number. On the other hand, there are 6 POS types supported by the set of the affixing rules. Three of these tags are common with Mi-POS tag set and these are noun, verb and adjective. Only the corresponding rules, i.e., the word relation rules and the affixing rules of the common tags will be considered to facilitate the comparison with the results of Mi-POS.

To build the POS tag dictionary to be used in the implementation of RPOS, we use the manually tagged training corpus to extract all unique words and all possible POS tags that could be associated with them. As the algorithm of RPOS states, before applying the disambiguation rules, words will be assigned a list of their possible POS tags which will be used in the word relation rules to narrow down the possibilities.

In our comparison we use the same two testing corpora to

evaluate the results of RPOS only on the tokens tagged with one of the following tags: noun, verb, adjective, adverb, preposition, conjunction and cardinal number. Only words that are given one tag by RPOS will be considered as correct. Other cases where RPOS fails to assign a single tag to a specific word, i.e., the disambiguation rules fail to keep only one POS tag for the word, will not be considered correct results for the RPOS tagger.

TABLE VII: TAGGING ACCURACY FOR NEWS ARTICLES FOR THE RPOS TAGGER BASED ON WORD'S CATEGORY

POS Type	All Tokens (A)	Correctly Tagged (B)	Accuracy $(\frac{B}{A} \times 100)\%$	Time (seconds)
Noun	1473	1155	78.41%	0.30
Verb	912	513	56.25%	0.27
Adjective	337	109	32.344%	0.28
Adverb	218	54	24.77%	0.27
Preposition	699	36	5.15%	0.27
Conjunction	271	7	2.58%	0.29
Cardinal Number	433	153	35.33%	0.27
Total	**4343**	**2027**	**46.67%**	

TABLE VIII: TAGGING ACCURACY FOR NON-NEWS ARTICLES FOR THE RPOS TAGGER BASED ON WORD'S CATEGORY

POS Type	All Tokens (A)	Correctly Tagged (B)	Accuracy $(\frac{B}{A} \times 100)\%$	Time (seconds)
Noun	3604	2276	63.15%	0.34
Verb	1856	764	41.16%	0.32
Adjective	522	118	22.6%	0.35
Adverb	283	47	16.61%	0.32
Preposition	974	35	3.59%	0.35
Conjunction	772	39	5.05%	0.35
Cardinal Number	369	104	28.18%	0.33
Total	**8380**	**3383**	**40.37%**	

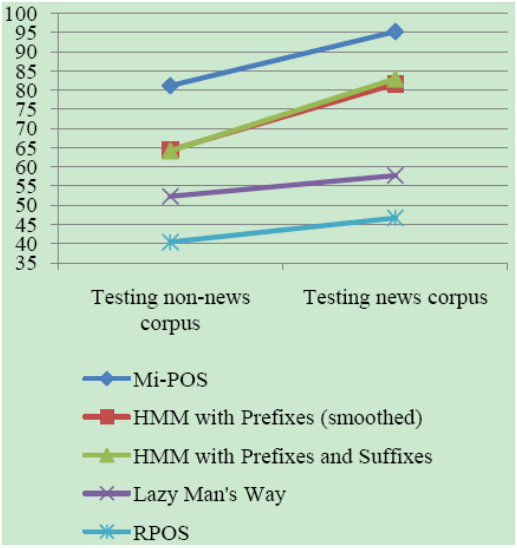


Fig. 3. The comparison of Mi-POS with the two used HMMs, the Lazy Man's Way and RPOS using the two testing corpora.

Table VII and Table VIII show the results of RPOS tagger using the news articles corpus and the non-news articles corpus respectively. All words in each sentence need to be tagged first then only words with a specific tag will be counted to calculate the individual accuracy for that tag.

Fig. 3 shows the performance for all models using the two testing corpora. It is noticed that Mi-POS has better accuracy for both testing corpora. Also Fig. 3 shows that the results for news articles are better than non-news articles for all models.

V. DISCUSSION

The results of Mi-POS show that the accuracy of tagging Bernama news archive is higher than the accuracy of tagging articles from fields other than news. This is mainly due to the fact that the training corpus used to train the model is also composed of articles taken from Bernama news archive and thus testing the model on articles from the same distribution as the training corpus will provide much accurate results since the model has been trained on similar words and similar contexts. However, for non-news articles, the model may find many unseen words in totally new contexts that the model is not trained to recognize with high accuracy. This can be solved by including articles from various domains in the training corpus used to train the model helps increasing the tagging accuracy of the model. Also increasing the size of the training corpus makes it possible for the model to correctly generalize to new unseen contexts.

Experiments from the previous section show that Mi-POS outperforms HMMs, the Lazy Man's and RPOS taggers with higher training and testing accuracies. Also it is noticed that for all used models the accuracy is higher when using the first corpus that includes news articles. This is due to the fact that these systems are trained using corpus with articles from the same distribution, i.e., Bernama news archive.

In terms of the processing time, Mi-POS is significantly faster than HMMs since fewer computations are required. Also the additional computations of the linear successive abstraction make the smoothed HMM with prefixes much slower than the other unsmoothed model with prefixes and suffixes.

When using the HMMs, and during both the training and testing phases, sentences are extracted from the articles and passed as a sequence of tokens to the model. In our experiments, if the sequence length exceeds a specific threshold of 13, the sentence will be segmented into two parts to reduce the complexity of the calculations. Each token in the sequence is first assigned all possible tags seen in the training corpus. Different tag sequences are then build for each sentence and the best tag sequence is chosen by the HMM as the most suitable POS tags for the sentence. The process of assigning all possible tags and building all tag paths for a sentence make the processing time of the HMMs much higher than that of the proposed model.

As previously mentioned, the Lazy Man's tagger tends to map Malay words to their tagged English words using a similarity formula. This mapping may be incorrect since the two languages have different structures. The nature of each language makes it difficult to map words since some phrases may be translated into a single word or vice versa. This is one

major source of the incorrect results of the Lazy Man's tagger. Furthermore, the results of the automatic Google Translator are not 100% accurate for some sentences. Also the Malay-English lexicon used to map Malay words to English words in the implementation of the tagger may not include all Malay words exist in the corpus which causes incorrect mappings and thus inaccurate tagged results.

The results of RPOS tagger show that the best achieved accuracy was for tagging nouns from both test corpora followed by tagging verbs. However, for the remaining five categories, the accuracy is relatively lower. This is due to the observation that most nouns and verbs where correctly tagged with only one tag as NN or VB respectively. Other words with other POS tags are tagged with multiple tags for the same word. Although RPOS processing time is very fast, it fails to assign a unique POS tag especially for the majority of the prepositions and conjunctions even if the right tag is selected as one of the possible tags for the word. Therefore, the total accuracy for all the seven categories of RPOS is low. A possible explanation for this is the limited number of word relation rules used in RPOS. It is also worth mentioning that changing the POS tag dictionary will certainly affect the tagging results. Therefore, the limitations of the POS tag dictionary in terms of tags supported and words number may be a possible cause of the low accuracy of the RPOS tagger.

Unlike other taggers, Mi-POS does not require any external resources such as a bilingual dictionary or an automatic translator. Also Mi-POS uses simple calculations based on the probabilistic methods to find the correct tags. Therefore, Mi-POS has shown higher accuracies and much faster processing time. The main error source is the limited training corpus used to train the model which makes the tagger biased to a specific text type (e.g., news articles).

VI. Conclusion

In this paper, we have presented a machine learning approach for the development of a Malay POS tagger called Mi-POS. The MaxEnt model used to develop the tagger is implemented using the open source NLP tool library OpenNLP. Mi-POS shows an accuracy of 95.16% on tagging corpus that is from the same type as the training corpus, and an accuracy of 81.12% when using a corpus from a different type. The proposed tagger also outperforms other Malay POS taggers with better accuracy and processing time.

We have shown the potentials of using statistical approaches to build a Malay POS tagger which out performs other rule-based taggers especially when using a large corpus for training. Also we have shown the limitations and drawbacks of existing Malay POS taggers applied using the conditions mentioned in the conducted experiments.

As future work, there are some improvements to be considered to enhance the performance of the system. A corpus of different types of articles other than news may be considered to train the model such as social media content, emails, and medical reports and so on in order to increase the accuracy of the system for these types. The variety and the size of the training data is an important factor to enhance the accuracy of the Mi-POS for assigning POS tags to unknown words correctly.

References

[1] T. Brants, "TnT: A statistical part-of-speech tagger," *Applied Natural Language Processing*, Stroudsburg, PA, USA, 2000, pp. 224-231.

[2] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Stroudsburg, PA, USA, 2003, pp. 173-180.

[3] S. Khoja, "APT: Arabic part-of-speech tagger," presented at the Student Workshop at NAACL, 2001.

[4] T. Solorio and Y. Liu, "Part-of-speech tagging for English-Spanish code-switched text," *Empirical Methods in Natural Language Processing (EMNLP)*, Honolulu, Hawaii, 2008.

[5] E. Brill, "A simple rule-based part of speech tagger," *Applied Natural Language Processing*, Stroudsburg, PA, USA, 1992, pp. 152-155.

[6] M. Hassan, N. Omar, and M. J. A. Aziz, "Statistical Malay part-of-speech (POS) tagger using Hidden Markov approach," *Semantic Technology and Information Retrieval (STAIR)*, Putrajaya, Malaysia, 2011, pp. 231-236.

[7] G. Knowles and Z. M. Don, *Word Class in Malay: A Corpus-Based Approach*, Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka, 2006.

[8] R. Alfred, A. Mujat, and J. H. Obit, "A ruled-based part of speech (RPOS) tagger for Malay text articles," *Intelligent Information and Database Systems*, Berlin, Heidelberg, 2013, pp. 50-59.

[9] D. B. D.Pustaka, *Tesaurus Bahasa Melayu Dewan*, Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka, 2005.

[10] N. Zamin, A. Oxley, Z. A. Bakar, and Y. A. Farhan, "A lazy man's way to part-of-speech tagging," *Knowledge Management and Acquisition for Intelligent Systems*, Berlin, Heidelberg, 2012, pp. 106-117.

[11] G. Knowles and Z. M. Don, "Tagging a corpus of Malay texts, and coping with 'syntactic drift'," in *Proc. the Corpus Linguistics 2003 Conference*, Lancaster, 2003, pp. 422-428.

[12] Z. M. Don, "Processing natural Malay texts: A data-driven approach," *TRAMES*, vol. 14, pp. 90-103, 2010.

[13] F. Pisceldo, M. Adriani, and R. Manurung, "Probabilistic part Of speech tagging for Bahasa Indonesia," presented at the Third International MALINDO Workshop, Colocated Event ACL-IJCNLP, Singapore, 2009.

[14] J. Baldridge. (2005) The opennlp project. [Online]. Available: https://opennlp.apache.org/

[15] BERNAMA. BERNAMA archived news. [Online]. Available: http://www.bernama.com/bernama/v8/newsarchive.php

[16] T. Baldwin and S. Awab, "Open source corpus analysis tools for Malay," in *Proc. the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006, pp. 2212-2215.

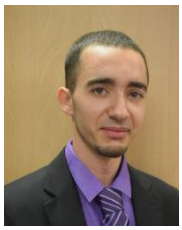
[17] Center for Sprog Teknologi. CST's Part of Speech Tagger. [Online]. Available: https://opennlp.apache.org/

[18] Malay / English Online Dictionary. [Online]. Available: https://opennlp.apache.org/

Benjamin Chu received his masters degree in information technology from Malaysia University of Science and Technology. He is currently pursuing his PhD from the University of Kiel, Germany. He is currently the staff researcher in MIMOS Berhad, specializing in the areas of natural language processing (NLP). He has authored several papers and invented over 15 intellectual properties related to NLP.

Mohamed Lubani received her masters degree in computer science (specialization in artificial intelligence) from Universiti Malaysia (UM). He is the research assistant in MIMOS Berhad working on the e-science project for Malay text understanding (a joint collaboration with Universiti Malaya). His focus areas in the project are entity recognition (ER) and word sense disambiguation (WSD).

Liew Kwei Ping received her bachelor degree in computer science from Universiti Sains Malaysia (USM). She is currently completing her masters in computer science from Universiti Malaya (UM). She is the research assistant in MIMOS Berhad working on the e-science project for Malay text understanding (a joint collaboration with Universiti Malaya). Her focus areas in the project are part-of-speech (pos) and relation extraction.



Khalil Bouzekri received his PhD in computer science (specialization in artificial intelligence/knowledge representation and reasoning) from the University of Montpellier, France.

He spent several years as sessional lecturer at the Institute of Technology of Montpellier and the University of Montpellier. He is currently the senior staff researcher in artificial intelligence applied research lab at MIMOS Berhad, where he heads the knowledge representation and reasoning and the natural language processing research teams. He has authored several research publications and invented over 20 intellectual properties.

Rohana Mahmud received her PhD in artificial intelligence from University of Manchester, UK.

She is the senior lecturer in the Department of Artificial Intelligence at Universiti Malaya (UM). Her areas of expertise are natural language (discourse structure, lexical relation, Malay language text processing), expert system (knowledge base system, multi-agent expert system, expert tutoring system), and education (AI in education, soft-skills, higher order thinking skills). She is currently a research collaborator in the e-science project on Malay text understanding with MIMOS berhad. She has authored several publications in these research areas.

Dr. Mahmud has obtained the International Exposition and Research Innovation Award in education, Sultan Idris Education University in 2013 and she was the program committee in International Workshop on Malay and Indonesian Language Engineering (MALINDO) Workshop (2011-2012) and LRE-Rel: Language Resources and Evaluation for Religious Texts Workshop (2012).

Dickson Lukose has obtained qualifications include B.Sc. (Hons) and PhD from Deakin University (Australia), and post-doctorate research as a leverhulme fellow at Loughborough University of Technology (UK). He has carried out teaching and research in artificial intelligence at Deakin University (Australia), Loughborough University of Technology (UK), University of Calgary (Canada) and University of New England (Australia).

He is the head of the knowledge technology research and development cluster at MIMOS Berhad. He is also the director of the artificial intelligence laboratory as well as the centre of excellence in semantic technologies. He has authored over 80 research papers and technical reports on the subject of artificial intelligence, knowledge acquisition and modeling for scientific journals and conference proceedings, chaired or co-chaired over 25 international conferences on these subjects, and edited a number of books in these areas.

Dr. Lukose spent many years working in financial services industry developing enterprise applications. He has done over 15 years of applied research in Artificial Intelligence, supported by research grants from Graphic Directions, Leverhulme Foundation, CSIRO, Australian Research Council, MIMOS and e-Science (MOSTI).