International Journal of Engineering & Technology, 7 (3.20) (2018) 6-14

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

POS-Taggging Malay Corpus: A Novel Approach Based on Maximum Entropy

[1,2]Juhaida Abu Bakar , [2]Khairuddin Omar , [2]Mohammad Faidzul Nasrudin and [2]Mohd Zamri Murah

[1]School of Computing, College of Arts & Sciences, Universiti Utara Malaysia, Malaysia
[2]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia
**Corresponding Author Email: juhaida.ab@uum.edu.my*

Abstract

Jawi and Roman scripts are represented Malay language. In the past, Jawi writings are widely used by the Malay community and foreigners; and it can be seen in the old documents. Old documents face the risk of background damage. In order to preserve this valuable information, there are significant needs to automated Jawi materials. Based on previous literature, POS-tags are known as the first phase in the automated text analysis; and the development of language technologies can barely initiate without this phase. We highlight the existing POS-tags approaches; and suggest the development of Malay Jawi POS-tags using extended ME-based approach on NUWT Corpus. Results have shown that the proposed model yielded a higher accuracy in comparison to the state-of-the-art model.

Keywords: *NLP pipeline task, POS-tags, tagging approach, Malay language, Jawi*.

1. Introduction

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages (Bird et al., 2009). In other hand, Information and Communication Technology (ICT) is used as an enabler to reduce the digital gap between the urban and rural community (Ali & Mohd Safar, 2011). ICT also used as a teaching and learning tool and can be used to increase productivity, efficiency and effectiveness of the management system. Jawi writing is a Malay writing with Arabic influences that have been used nearly 700 years ago. This is confirmed by the relevation of Terengganu Inscribed Stone, dated 1303 AD (Amat Juhari, 1991). In the past, these writings are widely used by the Malay community and foreigners who have diplomatic relationships, commerce, religious mission and such. At that time, the Malay language is the lingua franca of this region. So there are many Malay heritages such as manuscripts, religious books, letters, documents and other agreements in the Jawi scripts (Che Wan Shamsul Bahri, Khairuddin, Mohammad Faidzul, Mohd Zamri, & Azmi, 2013). However, old documents face the risk of background damage. Varying contrast, smudges, dirty background, ink through page, outdated paper and uneven background are an examples of background damage. The old Malay manuscripts which are a few hundred years of age are not legible even after preservation process by the library (Yahya et al., 2009). In order to preserve this valuable information, there are significant needs to do the first phase in the automated text of the Jawi text on the materials. Thus, by using ICT, research on POS-tags will help the effort.

Assigning syntactic categories to words is an important pre-processing step for most Natural Language Processing (NLP) applications (Biemann, 2010). Part-of-speech tagging or POS-tags is an important feature in NLP for word-category analysis. Effective analysis of Malay corpora can thus, be maximized through POS-tags; regardless of the writing system – the Roman (Rumi) or the Jawi script. It is generally accepted that the application of POS-tags in NLP applications can greatly improve the quality of NLP tasks. That being said, developing high-quality and fast tagging systems is still deemed to be a problem; despite the applications of several different POS-tags models and methods in various languages.

POS-tagging is the process of contextually assigning syntactic categories (noun, verb, etc.) with the most probable sequence to each word in a sentence. This task is a complex algorithmic process since one particular word might be associated with several possible tags. For example, the Malay word, "*menggembirakan*" (gloss: delightful) can be a verb (as in "*Sara menjalani kehidupan yang menggembirakan di China*") (gloss: Sara lived a happy life in China) or an adjective (as in "*Kejayaan Lim sungguh menggembirakan keluarganya*") (gloss: Lim's success makes his family happy). Malay adjectives can be easily

identified if the words are preceded by intensifiers such as "*amat*", "*sungguh*", "*sekali*", "*paling*" and "*agak*". Yet, it is the opposite in the case of non-adjectives; whereby, over 11% of the words in the hand-tagged Malay corpus are ambiguous (Hassan et al., 2011). Correspondingly, in recent years, there has been a growing interest in developing data-driven disambiguation applications.

POS-tags can be seen as a disambiguation task since the mapping between words and the tag-space is usually one-to-many (Zamora-Martinez et al., 2009). Two possible sources of information can be used to accurately predict the correct POS-tags of a word – contextual information and lexical information. The former is identified based on the different sequences of tags in a sentence. While some POS sequences are common; others are unlikely or impossible. For example, in Malay, prepositional phrases of direction is likely to be followed a verb, a preposition or a noun. On the other hand, the latter is identified based on the semantic value of word itself. For example, the word "ڤوكول (*pukul*)" (gloss: hit) can either be a verb or a noun. According to Knowles & Zuraidah (2003), the words needs to be analyzed through particular semantic rules to discover whether the meaning is, (first), to hit something, or (second), a special Malay adverbial used to specify the hour in indicating time. However, by utilizing a specific model of statistical and automated learning methods, features (sequences) of words can be listed without the needs to devise rules that are overcomplicated.

Different approaches have been proposed for the disambiguation tasks of POS-tags. The differences are based on either their internal model, the number of trainings or the information they need to process (Zamora-Martinez et al., 2009). In general, these different techniques can be categorized into three major categories: rule-based, statistical-based and transformation-based approaches.

The classical techniques (Rule-based approach) assign its corresponding POS-tags by employing certain lexicography rules. POS-tags which are designed using this approach consists of two stages of architecture (Jurafsky & Martin, 2009). The first stage involves extracting lexicographical data from the dictionary and assigning all probable POS-tags to every word match. The second stage involves employing handcrafted disambiguation rules in order to discover the most appropriate tag for each word.

In the case of automated tagging based on statistical information (statistical-based approach), a lot of different models have been developed and employed. POS tagging based on Hidden Markov Model (HMM) (Bar-Haim et al., 2008; Hasan et al., 2007; Hassan et al., 2011; Wicaksono & Purwarianti, 2010), Maximum Entropy (ME) (Hassan et al., 2015; Huang & Zhang, 2009; Malecha & Smith, 2010; Ratnaparkhi, 1999), Recurrent Neural Network (Zamora-Martinez et al., 2009), Conditional Random Field (Awasthi et al., 2006) and Support Vector Machine (Søgaard, 2010) are among the models under this category. They are designed based on the statistical occurrences of tag *n-grams* and *word-tag* frequencies which provide the information needed to identify the most probable tag sequence (Zamora-Martinez et al., 2009).

Transformation-based approaches combine both rule-based and statistical-based approaches. POS-tags based on transformation-based approaches (Brill, 1995) are designed to automatically derive the possible rules directly from the corpora. Recent years, researchers are moving forward to use machine learning approaches such as Deep learning (Boonkwan & Supnithi, 2017), Neural Network (Li et al., 2017; Viani et al., 2017) and optimization approach such as ant colony-based algorithm (Othmane, Fraj & Limam, 2017). Ant colony-based algorithms are among the most efficient methods to resolve optimization problems modeled as a graph. The collaboration of ants having various knowledge creates a collective intelligence and increases efficiency. This study shows, POS-tagging problems are considered as an optimization problem which modeled as a graph whose nodes correspond to all possible grammatical tags given by a morphological analyzer for words in a sentence and the goal is to find the best path (sequence of tags) in a graph. Both vocalized and non-vocalized texts used to performed experiments and tested two different tagsets containing fine and coarse grained composite tags (Othmane, Fraj & Limam, 2017).

Based on two previous studies in Malay studies (Hassan et al., 2015, 2011), the performances of POS tagging using HMM and ME models have been compared. HMM for Malay Roman script yielded 67.9% accuracy based on the morphological data gathered; and 94% with TnT. On the other hand, ME with *MaxEnt* and SVM with *SVMTools* reached the overall accuracies of 96% and 99.23% respectively (Hassan et al., 2015). In another similar study on Bahasa Indonesia (Pisceldo et al., 2009), an investigation on ME and CRF have been done. ME gives better results (in terms of accuracy) in comparison to CRF. ME recorded an accuracy of 97.57% while CRF recorded an accuracy of 91.15% for two tag sets containing 37 and 25 POS-tags (Pisceldo et al., 2009).

The objective of this study is to investigate and identify the most appropriate approach for the disambiguation tasks in Malay Jawi POS-tags. Our study is based on the specific contextual information which are related in the Jawi script of Malay corpora.

In Section 2, the standard probabilistic model for POS-tags is presented and discussed. In section 3, ME-based probabilistic model for Malay Jawi will be presented. In section 4, The NUWT Corpus which is used for training and testing POS tagger (Juhaida et al., 2016) is discussed. In section 5, the related contextual information for words and its neighbouring words in the Malay Jawi script will be discussed morphologically (in terms of their suffix/prefix features). In section 6, the training procedure and parameter-setting of ME-based probabilistic model is explained thoroughly. In section 7, the end results and comparative analysis with other methods are presented; and in Section 8, discussions and the conclusion on the findings will be remarked.

2. Pos-Tags Probabilistic Model

A probabilistic model employs POS-tags through the conditional probabilities given by the surrounding contextual features; whereby these probability values are obtained from a manually-tagged corpus (Zamora-Martinez et al., 2009). Let $T = \{t_1, t_2, \dots, t_{|T|}\}$ be a set of POS-tags and $\Omega = \{w_1, w_2, \dots, w_{|\Omega|}\}$ be the vocabulary of the application. The goal is to find the sequence of POS-tags that maximizes the probability associated to a sentence $w_1^n = w_1 w_2 \dots w_n$, i.e.:

$$\grave{t}_1^n = \underset{t_1^n}{argmax}\, P(t_1^n|w_1^n). \quad (1)$$

Using Bayes' theorem, the problem is reduced to:

$$\grave{t}_1^n = \underset{t_1^n}{argmax}\, P(w_1^n|t_1^n)P(t_1^n). \quad (2)$$

Estimating the values of these parameters can be time consuming since some levels of assumptions are needed – in order to simplify the computational process of the expression (Merialdo, 1994; Zamora-Martinez et al., 2009). For these models, it is assumed that words are independent of each other; and a word's identity only depends on its tag. Correspondingly, we would be able to obtain this *lexical* probability,

$$P(w_1^n|t_1^n) \approx \prod_{i=1}^n P(w_i|t_i). \quad (3)$$

Another assumption establishes that the probability of one tag to appear only depends on its predecessor tags,

$$P(t_1^n) \approx \prod_{i=1}^n P(t_i|t_{i-1}, t_{i-2}, \dots, t_{i-k+1}), \quad (4)$$

if a *k*-gram class is able to obtain the *contextual* probabilities.

With these assumptions, a typical probabilistic model following equations (2), (3) and (4) is expressed as follows:

$$\hat{t}_1^n \approx \underset{t_1^n}{argmax} \prod_{i=1}^n P(w_i|t_i)P(t_{i-1}, t_{i-2}, \dots, t_{i-k+1}). \quad (5)$$

whereby $\hat{t}_1^n$ is the best estimation of POS-tags for the given sentence w_1^n. Nonetheless, two limitations on the probabilistic model are identifiable: (1) it does not model long-distance lexical relationships, (2) the contextual information takes into account the context on the left while the context on the right is not considered (Zamora-Martinez et al., 2009).

3. Maximum Entropy Model for Malay Jawi Pos-Tags

Maximum Entropy (ME) belongs to the family of classifiers known as the exponential or log-linear classifiers (Jurafsky & Martin, 2009). ME is designed to work by extracting some set of features from the input, combining them linearly (multiplying each by a particular weight and then add them up), and then, using this sum as an exponent (Jurafsky & Martin, 2009). This method allows high flexibility in utilizing contextual information and assigns an appropriate tag based on a probability distribution. The probability distribution should have the highest entropy values found on the training corpus and it must be in accordance to certain conditional values. Correspondingly, ME models the POS-tags task as:

$$p(h,t) = \pi\mu \prod_{j=1}^k \alpha_j^{f_j(h,t)} \quad (6)$$

where *h* is a 'history' of observation and tag sequences, *t* is a tag, μ is a normalization constant, $f_j(h,t)$ is the feature functions with binary values of 0 and 1, and μ and $\alpha_1, \dots, a_k$ are model parameters (Pisceldo et al., 2009).

The model parameters must be set in a specific value in order to maximize the entropy of the probability distribution; and additionally, the entropy is subjected to the constraints imposed by the value of the f_j feature functions from the training data (Pisceldo et al., 2009). The Generalized Iterative Scaling (GIS) algorithm, Improved Iterative Scaling (IIS) and the optimized version *Megam* commonly trained these parameters. It is used to suit the log-linear model (Nurwidyantoro & Winarko, 2012). According to (Pisceldo et al., 2009; Ratnaparkhi, 1996), the underlying philosophy is to choose the model that makes the fewest assumptions about the data whilst still remaining consistent with it.

4. Nuwt Corpus

The NUWT corpora sources were gathered from three different genres of documents – standard written and conversational Malay, Malay narratives and Malay translation of *Quranic* verses. The NUWT corpora are written using Jawi-specific-Buckwalter code (Juhaida et al., 2013). The first source is an annotated corpus named the "Malay corpus". It contains 21 tags and 18,135 tokens with 1,381 words that have ambiguous tags. The corpus was originally prepared by (Hassan et al., 2011) using the Dewan Bahasa dan Pustaka (DBP) tag set; and was written using the Roman script. The second source is a grammatical corpus named the "Malay corpus UKM-DBP"; and is a collection of story books with 12,304 words. The corpus was developed (Nurul Huda et al., 2012) according to the DBP tag set and was also originally written in the Roman script. It has five main tags – with respective elaborated sections for each main tag (Juhaida et al., 2016). The third source is from the "Quranic Malay written in Jawi character Corpus" (Suliana et al., 2011) which is an unannotated text of Quranic translations. It contains a collection of 114 chapters with 157,388 words. The corpus is written in Jawi standard Unicode (UTF8).

5. Contextual Information

The training corpus was partitioned into ten parts of equal size. Fundamentally, the words that appeared in each partition played the functions of a testing corpus - enabling the dimensions of the feature sets to be reduced. Additionally, the technique of 10-fold cross-validation was used with 9 different models. From the cross-validation technique, the contextual information can thus, be extracted and concluded with the calculation of an average accuracy. Table 1 shows number of tokens for each fold.

Table 1 Number of tokens in training corpus for 9 models

Model	Number of tokens	
	Malay corpus	Malay corpus UKM-DBP
1	16,322	11,815
2	14,508	10,502
3	12,695	9,188
4	10,881	7,877
5	9,068	6,564
6	7,254	5,251
7	5,441	3,939
8	3,627	2,626
9	1,814	1,313

Based on the previous work (Hassan et al., 2011) and Jawi rules (Hamdan, 1999; Ismail, 1991), we consider several types of features, which is likely suitable for Malay Jawi.

Affix Features

Affix features are the simplest type of features in Malay language. According to (Suliana et al., 2011), for Malay language, a derived word can be described as a combination of a prefix, a circumfix, a suffix or an infix with a root word. Table 2 exemplifies the differences between the spelling rules for the suffix "+an" in the Roman and the Jawi scripts respectively.

Table 2 The Roman and Jawi spelling rules for suffixes

Jawi	Roman Scripts
+أن	+an
+ن	+an
+ءن	+an
+ان	+an

Source: Suliana et al. (2011)

These features are likely to be most useful in languages that utilize morphological rules to modify word structures and meanings such as the Malay language. Additionally, the features have been automatically constructed from the training corpus by recording all prefixes and suffixes up to a certain length. Table 3 shows the affixation rules applied in the context of Jawi script. From the table, the valid length of affixation in the Jawi script for Malay language is up to 4 morphemes on either side of the stem.

Table 3 Derivative Jawi writing for prefixes and suffixes

Jawi	Roman Scripts	Jawi	Roman Scripts	Jawi	Roman Scripts
+انتي	*anti+*	+م	*me+*	+فر	*per+*
+اءوتو	*auto+*	+مم	*mem+*	+فولي	*poli+*
+ب	*be+*	+من	*men+*	+فرا	*pra+*
+بل	*bel+*	+مڠ	*meng+*	+فرو	*pro+*
+بر	*ber+*	+مڠ	*menge+*	+س	*se+*
+بي	*bi+*	+ممڤر	*memper+*	+سوب	*sub+*
+د	*di+*	+ڤنچا	*panca+*	+سوڤرا	*supra+*
+دڤر	*diper+*	+ڤ	*pe+*	+سوا	*swa+*
+دوي	*dwi+*	+ڤل	*pel+*	+تاتا	*tata+*
+ايكا	*eka+*	+ڤم	*pem+*	+ت	*te+*
+جورو	*juru+*	+ڤن	*pen+*	+تر	*ter+*
+ك	*ke+*	+ڤڠ	*peng+*	+تري	*tri+*
+مها	*maha+*	+ڤڠ	*penge+*	+تونا	*tuna+*
+ه	*+ah*	+يسمي	*+isme*	+ون	*+wan*
+ات	*+at*	+كن	*+kan*	+واتي	*+wati*
+يه	*+iah*	+من	*+man*	+وي	*+wi*
+ين	*+in*	+نيتا	*+nita*		

Source: Hamdan (1999)

Neighbourhood Features

In addition to using the current words, the tags of surrounding words can also be used as features (Malecha & Smith, 2010). In this section, we contrastingly used the tags of surrounding words as features. A common example from the Malay linguistic rule is that the word following a cardinal number is often a noun or a verb; but infrequently, it can also be followed by an adjective or preposition. We expect these features to be beneficial to the process of classification in languages that heavily use modifiers and word positioning.

5. Setting the Me-Based Pos Taggers

Learning ME model can be done via a generalization of the logistic regression learning algorithms. We want to find the parameter, *w*, which maximizes the likelihood of *M* training samples:

$$\hat{w} = \underset{w}{argmax} \prod_i^M P\left(y^{(i)}|x^{(i)}\right) \quad (7)$$

Experiments on ME-based model is more complex than any other models. Two major experiments conducted which are identifying parameters and features to select the best model. Five experiments have been conducted to identify three parameters that can maximize classifying accuracy. Three forms of basic parameters used in the Natural Language ToolKit (NLTK) module are algorithms, log likelihood delta, and training iteration. Table 4 shows the default parameters in NLTK. Table 5 shows the parameters used in second, third, forth and fifth experiment with the aptitude to select the best parameters that can maximize accuracy.

Table 4 NLTK default parameters

Parameters	Type
Algorithm	IIS
Log likelihood delta	Stop when the repetition log likelihood fixed less than the previous log likelihood
Maximum iteration	100

Table 5 NLTK parameters for subsequent experiments

Parameters	Experiment			
	2	3	4	5
Algorithm	IIS	GIS	GIS	GIS
Log likelihood delta	0.01	0.01	0.01	0.005
Maximum iteration	10	10	20	20

Five experiments on features have been conducted using NLTK module - providing the finest parameters with the aptitude to select the best features that can maximize accuracy. The set of features developed are shown in Table 6 and 7. The experiments listed are from a series of preliminary experiments which aims to determine the number of adjustments needed to provide the highest accuracy.

Table 6 Feature setting for the five experiments

Feature Name	Experiment				
	1	2	3	4	5
suf-1-x	√				
suf-2-xx	√	√	√	√	√
suf-3-xxx	√	√	√	√	√
pre-2-xx			√	√	√
pre-3-xxx		√	√	√	√
word-*w*₀-1	√	√	√		
tag-*T*₀-1				√	√
tag-*T*₀-2					√

Table 7 Description of the feature set

Feature Name	Condition	Meaning
suf-1-x	= word[-1:] == 'x'	Word ends with "x"
suf-2-xx	= word[-2:] == 'xx'	Word ends with "xx"
suf-3-xxx	= word[-3:] == 'xxx'	Word ends with "xxx"
pre-2-xx	= word[0:1] == 'xx'	Word begins with "xx"
pre-3-xxx	= word[0:2] == 'xxx'	Word begins with "xxx"
word-*w*₀-1	= word[-1] == *w*	Previous word is *w*
tag-*T*₀-1	= tag [-1] == *T*	Previous word has tag *T*
tag-*T*₀-2	= tag [-2] == *T*	Two word back has tag *T*

Source: Malecha & Smith (2010)

In subsequent experiments, only prefixes and suffixes with 2 and 3 morphemes are taken into considerations. The average accuracy was found to have increased with the features as displayed in the results section of these experiments.

6. Me-Based Tagger Performance

In determining the best features of the corpus, the set features were run to identify the features with the highest average accuracy in three phases. In the first phase, NLTK default parameters were used in producing these results. Correspondingly, five experimental results using five sets of features are presented in Tables 8 and 9 for Malay corpus and Malay corpus UKM-DBP respectively. By using different feature settings, it displays that the third experiment gave higher level of average accuracy on the Malay corpus. Meanwhile, feature setting for the fourth experiment gives higher average accuracy on Malay corpus UKM-DBP.

Table 8 Experimental results for Malay Corpus

Model (Epoch)	Fold	Malay Corpus				
		1*	2*	3*	4*	5*
1	90:10	88.20%	93.83%	94.27%	94.43%	92.17%
2	80:20	85.86%	92.50%	92.48%	92.42%	92.14%
3	70:30	83.22%	90.24%	91.38%	90.87%	90.81%
4	60:40	82.81%	90.57%	91.90%	91.80%	91.90%
5	50:50	82.11%	89.93%	91.01%	90.75%	90.83%
6	40:60	80.81%	88.92%	90.06%	89.92%	90.11%
7	30:70	79.56%	87.01%	88.32%	88.06%	88.22%
8	20:80	76.23%	82.92%	84.30%	84.27%	84.35%
9	10:90	70.94%	77.69%	79.94%	80.61%	81.20%
	Average	**81.08%**	**88.18%**	**89.30%**	**89.24%**	**89.08%**

* based on experiment setting

Table 9 Experimental results for Malay Corpus UKM-DBP

Model (Epoch)	Fold	Malay Corpus UKM-DBP				
		1*	2*	3*	4*	5*
1	90:10	60.26%	66.12%	65.79%	65.62%	65.38%
2	80:20	59.85%	64.80%	65.99%	66.08%	65.42%
3	70:30	61.28%	65.16%	66.03%	66.03%	65.79%
4	60:40	59.20%	63.66%	64.28%	64.25%	63.16%
5	50:50	59.14%	64.05%	64.96%	65.27%	62.76%
6	40:60	56.12%	60.70%	61.89%	62.15%	60.09%
7	30:70	55.81%	60.60%	61.48%	62.04%	62.48%
8	20:80	52.25%	55.74%	57.08%	56.67%	57.47%
9	10:90	52.55%	54.65%	56.06%	56.05%	56.45%
	Average	**57.39%**	**61.72%**	**62.62%**	**62.69%**	**62.11%**

* based on experiment setting

In the second phase, the set NLTK parameters were run to identify the highest average accuracy. Correspondingly, five experimental results using five set NLTK parameters are presented in Tables 10 and 11 for Malay corpus and Malay corpus UKM-DBP respectively. Experiment 1 shows the best average accuracy achieved using NLTK default parameters. By using different parameter settings (refer Table 5), it displays that the third experiment gave higher level of average accuracy on the Malay corpus. Meanwhile, parameters setting for the fourth experiment give higher average accuracy on Malay corpus UKM-DBP.

Table 10 Experimental results for various parameters in Malay Corpus

Model (Epoch)	Fold	Malay Corpus				
		1*	2*	3*	4*	5*
1	90:10	94.27%	94.38%	94.10%	94.10%	94.05%
2	80:20	92.48%	92.31%	92.36%	92.36%	92.20%
3	70:30	91.38%	91.51%	91.20%	91.20%	91.12%
4	60:40	91.90%	91.74%	91.94%	91.94%	91.80%
5	50:50	91.01%	90.88%	91.04%	91.04%	91.01%
6	40:60	90.06%	90.39%	90.45%	90.45%	90.42%
7	30:70	88.32%	88.55%	88.86%	88.86%	88.72%
8	20:80	84.30%	84.89%	85.03%	85.03%	84.92%
9	10:90	79.94%	80.28%	80.79%	80.79%	80.78%
	Average	**89.30%**	**89.44%**	**89.53%**	**89.53%**	**89.45%**

* based on experiment setting

Table 11 Experimental results for various parameters in Malay Corpus UKM-DBP

Model (Epoch)	Fold	Malay Corpus UKM-DBP				
		1*	2*	3*	4*	5*
1	90:10	65.62%	65.95%	66.69%	66.61%	66.45%
2	80:20	66.08%	66.03%	66.45%	66.45%	66.03%
3	70:30	66.03%	66.69%	66.86%	66.91%	66.64%
4	60:40	64.25%	64.87%	64.83%	64.89%	64.65%
5	50:50	65.27%	65.58%	66.18%	66.09%	65.86%
6	40:60	62.15%	62.72%	62.42%	62.51%	62.28%
7	30:70	62.04%	62.54%	62.09%	62.09%	62.01%
8	20:80	56.67%	57.48%	57.38%	57.38%	57.25%
9	10:90	56.05%	56.63%	56.55%	56.55%	56.51%
	Average	**62.69%**	**63.17%**	**63.27%**	**63.28%**	**63.07%**

* based on experiment setting

In the third phase, re-assessment has been made to the set features to refine results achieved by using NLTK default parameters in phase one. It is proven that in the second phase, the best parameter is different from the first phase. Five experimental results with best parameter achieved in second phase using five sets of features are presented in Tables 12 and 13 for Malay corpus and Malay corpus UKM-DBP. It displays that the fifth experiment gave higher level of average accuracy on the Malay corpus and Malay corpus UKM-DBP. A consensus results obtained from these two different types of corpus although these two use different genre. It shows that the Malay corpus and Malay corpus UKM-DBP used the same feature sets to excel the highest average accuracy.

Table 12 Experimental results for re-assessment features in Malay Corpus

Model (Epoch)	Fold	Malay Corpus				
		1*	2*	3*	4*	5*
1	90:10	88.48%	93.61%	94.10%	94.16%	94.16%
2	80:20	85.80%	92.23%	92.36%	92.67%	92.53%
3	70:30	83.83%	90.54%	91.20%	91.68%	91.40%
4	60:40	83.39%	90.77%	91.94%	92.06%	92.05%
5	50:50	83.20%	90.11%	91.04%	90.90%	90.84%
6	40:60	82.00%	89.13%	90.45%	90.16%	90.23%
7	30:70	80.38%	87.34%	88.86%	89.03%	88.94%
8	20:80	77.34%	83.33%	85.03%	84.72%	84.96%
9	10:90	72.47%	77.88%	80.79%	80.72%	81.29%
	Average	**81.88%**	**88.33%**	**89.53%**	**89.57%**	**89.60%**

* based on experiment setting

Table 13 Experimental results for re-assessment features in Malay Corpus UKM-DBP

Model (Epoch)	Fold	Malay Corpus UKM-DBP				
		1*	2*	3*	4*	5*
1	90:10	61.17%	65.54%	65.79%	66.61%	66.36%
2	80:20	61.46%	66.03%	66.36%	66.45%	66.61%
3	70:30	62.27%	66.25%	66.45%	66.91%	66.89%
4	60:40	60.42%	64.58%	65.22%	64.89%	65.02%
5	50:50	60.03%	64.99%	65.98%	66.09%	65.98%
6	40:60	56.91%	61.33%	62.20%	62.51%	62.48%
7	30:70	56.70%	60.81%	62.04%	62.09%	62.58%
8	20:80	53.01%	56.15%	57.43%	57.38%	57.73%
9	10:90	52.86%	54.92%	56.30%	56.55%	56.89%
	Average	**58.31%**	**62.29%**	**63.09%**	**63.28%**	**63.39%**

* based on experiment setting

For comparative and validation purposes, we tested our corpora using the standard HMM probabilistic model. Table 14 shows the contrasts between these two models. Ultimately, this study has identified that ME model provides a higher average accuracy compared to HMM model for both Malay Jawi corpora. It is possible that these results are due to its source of information to accurately predict the correct POS-tags of a word. A graph for the Malay corpus and the Malay corpus UKM-DBP is then plotted in Figure 1 to compare the values from the two corpora. It is highly probable that the lower level of accuracy in the Malay Corpus UKM-DBP is due to its genre – narratives.

Table 14 Results of the ME highest accuracy and comparison to HMM

Model (Epoch)	**Fold**	**Malay Corpus**		**Malay Corpus UKM-DBP**	
		ME	HMM	ME	HMM
1	90:10	94.16%	92.41%	66.36%	62.45%
2	80:20	92.53%	90.15%	66.61%	62.71%
3	70:30	91.40%	89.48%	66.89%	62.49%
4	60:40	92.05%	90.30%	65.02%	60.34%
5	50:50	90.84%	88.21%	65.98%	62.36%
6	40:60	90.23%	87.80%	62.48%	56.71%
7	30:70	88.94%	87.22%	62.58%	56.19%
8	20:80	84.96%	76.65%	57.73%	54.23%
9	10:90	81.29%	68.21%	56.89%	49.33%
	Average	**89.60%**	**85.60%**	**63.39%**	**58.53%**

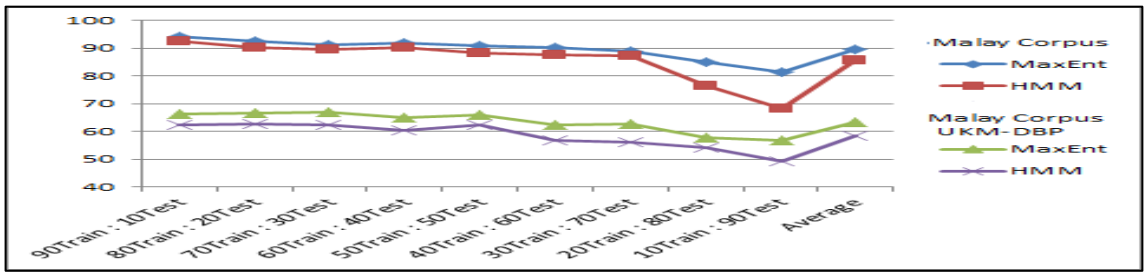

Figure 1 Graph between HMM and ME for Jawi Tagger

7.Discussion

This study focuses on evaluating the ME model for the development of POS Tags in NLP applications – primarily focusing on its application in the Jawi script of Malay language. The results show that the ME-based model is suitable to be applied to the Malay Jawi script due to its good analytical features on contextual information. Results have also shown that the ME-based model yielded higher accuracy level in comparison to the HMM probabilistic model. The lower level of accuracy in the Malay Corpus UKM-DBP is most probably due to the genre of the corpus.

Based on these findings, a probabilistic model (ME) that can categorize the Jawi-written Malay words into its accurate POS has been identified. For future research endeavours, other Jawi corpora such as the third corpus of NUWT Corpus shall be analyzed for greater reliability and validity. Correspondingly, other derivational words formed through other types of Malay affixations such as circumfix and infix can be added to be part of our future study in NLP applications on the Jawi script of Malay language. Production of the Jawi tagger using ME-based approach will be able to help the intermediate process on NLP onwards.

Acknowledgment

First author is fully supported by the Universiti Utara Malaysia scholarships. The material is based upon work supported by the Universiti Kebangsaan Malaysia (UKM) under Grant No. ERGS/1/2013/ICT1/UKM/3/5.

References

[1] Ali, S., & Mohd Safar, H. (2011). Internet usage in a Malaysian sub-urban community: A study of diffusion of ICT innovation. The Innovation Journal: The Public Sector Innovation Journal, 16(2), Article 6.

[2] Amat Juhari, M. (1991). Sejarah tulisan Jawi. Jurnal Dewan Bahasa, 35(11), 1001–1012.

[3] Awasthi, P., Rao, D., & Ravindran, B. (2006). Part Of Speech Tagging and Chunking with HMM and CRF. In Proceedings of NLPAI contest workshop during NWAI '06 (pp. 1–4). SIGAI Mumbai. Retrieved from http://publications.cse.iitm.ac.in/157/

[4] Bar-Haim, R., Sima'An, K., & Winter, Y. (2008). Part-of-speech tagging of Modern Hebrew text. Natural Language Engineering, 14(02), 223–251. doi:10.1017/S135132490700455X

[5] Biemann, C. (2010). Unsupervised Part-of-Speech Tagging in the Large. Research on Language and Computation, 7(2-4), 101–135. doi:10.1007/s11168-010-9067-9

[6] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python (1st ed.). USA: O'Reilly Media, Inc.

[7] Boonkwan, P., & Supnithi, T. (2017). Bidirectional Deep Learning of Context Representation for Joint Word Segmentation and POS Tagging. In International Conference on Computer Science, Applied Mathematics and Applications (pp. 184–196). Berlin: Springer.

[8] Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing : A Case Study in Part-of-Speech Tagging. Computational Linguistics, 21(4), 543–565.

[9] Che Wan Shamsul Bahri, C. W. A., Khairuddin, O., Mohammad Faidzul, N., Mohd Zamri, M., & Azmi, S. M. (2013). Machine Transliteration for Old Malay Manuscript. In 2nd International Conference on Machine Learning and Computer Science (IMLCS'2013) (pp. 23–26). Kuala Lumpur.

[10] Hamdan, A. R. (1999). Panduan menulis dan mengeja Jawi. Kuala Lumpur: Dewan Bahasa dan Pustaka.

[11] Hasan, F. M., UzZaman, N., & Khan, M. (2007). Comparison of Different POS Tagging Techniques (N-gram , HMM and Brill ' s tagger) for Bangla. In K.Elleithy (Ed.), Advances and Innovations in Systems, Computing Sciences and Software Engineering (pp. 121–126). Springer.

[12] Hassan, M., Nazlia, O., & Mohd Juzaiddin, A. A. (2015). Malay Part of Speech Tagger : A Comparative Study on Tagging Tools. Asia-Pacific Journal of Information Technology and Multimedia, 4(1), 11–23.

[13] Hassan, M., Nazlia, O., & Mohd Juzaidin, A. A. (2011). Statistical Malay Part-of-Speech (POS) Tagger using Hidden Markov Approach. In 2011 International Conference on Semantic Technology and Information Retrieval (pp. 231–236). IEEE.

[14] Huang, H., & Zhang, X. (2009). Part-of-speech tagger based on maximum entropy model. 2009 2nd IEEE International Conference on Computer Science and Information Technology, 26–29. doi:10.1109/ICCSIT.2009.5234787

[15] Ismail, D. (1991). Pedoman Ejaan Jawi yang Disempurnakan. Kuala Lumpur: Dewan Bahasa dan Pustaka.

[16] Juhaida, A. B., Khairuddin, O., Mohammad Faidzul, N., & Mohd Zamri, M. (2016). NUWT : Jawi-specific Buckwalter Corpus for Malay Word Tokenization. Journal of Communication and Information Technology, 15, 1–25.

[17] uhaida, A. B., Khairuddin, O., Mohammad Faidzul, N., Mohd Zamri, M., & Che Wan Shamsul Bahri, C. W. A. (2013). Implementation of Buckwalter transliteration to Malay corpora. In 2013 13th International Conference on Intelligent Systems Design and Applications (pp. 213–218). Serdang. doi:10.1109/ISDA.2013.6920737

[18] Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Second Edi.). New Jersey, USA: Pearson Education, Inc.

[19] Knowles, G., & Zuraidah, M. D. (2003). Tagging a corpus of Malay texts, and coping with "syntactic drift." In Proceedings of the corpus linguistics (pp. 422–428). Retrieved from http://eprints.lancs.ac.uk/8620/

[20] Li, Z., Chao, J., Zhang, M., Chen, W., Zhang, M., & Fu, G. (2017). Coupled POS Tagging on Heterogeneous Annotations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(3), 557–571.

[21] Malecha, G., & Smith, I. (2010). Maximum Entropy Part-of-Speech Tagging in NLTK (pp. 1–10). unpublished course-related report.

[22] Merialdo, B. (1994). Tagging English text with a probabilistic model. Computational Linguistics, 20(2), 155–172.

[23] Nurul Huda, M. S., Juhaida, A. B., Rafidah, A. K., Nurbaiti, T., & Khalijah, M. N. (2012). Pembangunan korpus cerpen bertag Bahasa Melayu: Analisis Linguistik Korpora. In Research, Invention, Innovation & Design (RIID 2012) (pp. 1–5). Universiti Teknologi MARA Kampus Melaka.

[24] Nurwidyantoro, A., & Winarko, E. (2012). Parallelization of Maximum Entropy POS Tagging for Bahasa Indonesia with MapReduce. International Journal of Computer Science Issues (IJCSI), Vol. 9(Issue 4), 1–6.

[25] Othmane, C. Z. B., Fraj, F. B., & Limam, I. (2017). POS-tagging arabic texts: A novel approach based on ant colony. Natural Language Engineering, 23(3), 419–439.

[26] Pisceldo, F., Adriani, M., & Manurung, R. (2009). Probabilistic Part Of Speech Tagging for Bahasa Indonesia. In Third International MALINDO Workshop, colocated event ACL-IJCNLP (pp. 1–6). Singapore.

[27] Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In Proceedings of the conference on empirical methods in natural language processing (pp. 133–142).

[28] Ratnaparkhi, A. (1999). Learning to Parse Natural Language with Maximum Entropy Models. Machine Learning, 34(1-3), 151–175. doi:10.1023/A:1007502103375

[29] Søgaard, A. (2010). Simple semi-supervised training of part-of-speech taggers. In Proceedings of the ACL 2010 Conference Short Papers (pp. 205–208). Uppsala: Association for Computational Linguistics.

[30] Suliana, S., Khairuddin, O., Nazlia, O., Mohd Zamri, M., & Hamdan, A. R. (2011). A Malay Stemmers for Jawi Characters. In D. Wang & M. Reynolds (Eds.), AI 2011: Advances in Artificial Intelligence (pp. 668–676). Perth, Australia: Springer Berlin / Heidelberg. doi:10.1007/978-3-642-25832-9_68

[31] Viani, N., Miller, T. A., Dligach, D., Bethard, S., Napolitano, C., Priori, S. G., Bellazzi, R., Sacchi, L., & Savova, G. K. (2017). Recurrent Neural Network Architectures for Event Extraction from Italian Medical Reports. In Conference on Artificial Intelligence in Medicine in Europe (pp. 198–202). Vienna: Springer.

[32] Wicaksono, A. F., & Purwarianti, A. (2010). HMM Based Part-of-speech Tagger for Bahasa Indonesia. In The 4th International MALINDO (Malay and Indonesian Language) Workshop (pp. 1–7).

[33] Yahya, S. R., Abdullah, S. S., Omar, K., Zakaria, M. S., & Liong, C. Y. (2009). Review on image enhancement methods of old manuscript with the damaged background. In Proceedings of International Conference on Electrical Engineering and Informatics (pp. 62–67). Bangi: IEEE.

[34] Zamora-Martinez, F., Castro-Bleda, M. J., Espana-Boquera, S., & Tortajada-Velert, S. (2009). Adding Morphological Information to a Connectionist Part-Of-Speech Tagger. In Current Topics in Artificial Intelligence (pp. 191–200). Seville: Springer Berlin Heidelberg.