PENALIZING UNKNOWN WORDS' EMISSIONS IN HMM POS TAGGER BASED ON MALAY AFFIX MORPHEMES

Hassan Mohamed[1], Nazlia Omar[2], Mohd Juzaiddin Ab Aziz[3]
Zuraini Zainol[4], Syahaneim Marzukhi[5].

[1,4,5] Cyber Security Centre,
Department of Computer Science,
Univerisiti Pertahanan Nasional Malaysia.

Email: [1]hassan@upnm.edu.my,
[4]zuraini@upnm.edu.my,
[5]syahaneim@upnm.edu.my

[2,3]Knowledge Tech. Group, Centre for AI Technology (CAIT),
Faculty of Information Science & Technology,
Universiti Kebangsaan Malaysia.

Email: [2]nazlia@ukm.edu.my,
[3]juzaiddin@ukm.edu.my

ABSTRACT

The challenge in unsupervised Hidden Markov Model (HMM) training for a POS tagger is that the training depends on an untagged corpus; the only supervised data limiting possible tagging of words is a dictionary. Any published dictionary can be used even if it does not include all the words found in a corpus. Therefore, training cannot properly map possible tags. The exact morphemes of prefixes, suffixes and circumfixes in the agglutinative Malay language was examined to assign unknown words' probable tags based on linguistically meaningful affixes using a morpheme-based POS guessing algorithm for baseline tagging. The tagger was first examined using HMM-Viterbi with unknown words handled by character-based prediction; next, it was examined using HMM-Viterbi with unknown words handled by morpheme-based POS guessing; lastly, it was examined using a combination of the first and second. HMM-Viterbi tagging with morpheme-based POS guessing was found to be better than HMM-Viterbi tagging with character-based prediction, especially at tagging unknown words not identified in the dictionary. The best combination proved to be using morpheme-based POS guessing with unknown word emissions replaced by a value proportionate to the marginal distribution of tags and words' ending information, given a maximum predefined length of six characters and ignoring affixed words in smoothing. This method proved to be satisfactory for guessing tags of words not in the dictionary, and it outperformed the baseline.

Keywords: *Malay POS tagger, morpheme-based, HMM*

1.0 INTRODUCTION

The Malay language is categorised as an agglutinative or derivative language where most of the words are formed by merging affixes with root words (Nik Safiah, 2010; Abdullah, 2006). Affixation is done in a way in which the affix is either added at the beginning (prefixes), the middle (infixes), the end (suffix) or at both beginning and end (circumfixes) of a root word. Due to the well-defined affixation rules, the word class of Malay derivative words can be intuitively guessed. Therefore, this study empirically examined the effectiveness of using Malay affix morphemes for handling unknown words in the unsupervised Hidden Markov Model (HMM) POS tagging. The idea of using the unsupervised approach in contrast to the supervised one was to avoid the provision of a large annotated corpus, which would be labour intensive, time consuming and high in costs, especially for under-resourced languages such as Malay.

The only supervision given in the traditional unsupervised HMM training is a dictionary. The dictionary matches the possible parts of speech to words in the training corpus. Therefore, it limits the tags of a word when assigning the HMM initial emissions before the training starts. In the traditional unsupervised HMM training, dictionaries are built from a tagged corpus (Banko *et al.*, 2004; Kupiec, 1992; Merialdo, 1994).

Therefore, this dictionary was attributed to an artificial one since it was a filtered dictionary where all words, with only tags that had certain relative frequencies, occurred in the annotated corpus that were taken as entries. This provided a scenario where every word in the corpus was considered as known words. In contrast to this scenario, for under-resourced languages such as Malay, a tagged corpus is not freely available. Therefore, a similar dictionary was unable to form; yet, some printed dictionaries were available and used as an alternative. However, the printed dictionary does not include most of the words that exist in the corpus, especially passive verbs and derivative words. This caused such words to be considered as unknown since the words that even appear in the corpus have unknown tags.

2.0 RELATED WORK

Several researchers had conducted their efforts to train unsupervised HMM POS tagger, which catered words not listed in the dictionary or ambiguous words. They exploited the words' ending with specified length to enlarge the training dictionaries (Miller *et al*, 2007; Ravi *et al.*, 2009); or directly estimated the initial emissions for unknown words (Cucerzan *et al.*, 2000; Garrette *et al.*, 2012); or directly estimated the lexical probabilities for ambiguous words (Goldberg *et al.*, 2008). There are also researchers who automatically built an annotated corpus and then trained the HMM using the supervised approach (Garrette *et al.*, 2013).

In order to enlarge the training dictionary, completely unknown lexicon w was created by exploiting the words' ending patterns as the exemplary, which was extracted from an example corpus, such as the WSJ corpus (Miller *et al.*, 2007). For each unknown lexical w, all immediate neighbours (words similar to w from the annotated corpus) were searched. The average of the exemplary lexical probability was used to obtain the lexical probability of w; i.e. $P(t|w)$. This method was very effective; even the application domain was significantly different from the exemplary text. The other method was to enlarge the training dictionary. The unknown words with their predicted POS were added as entries (Ravi *et al.*, 2009). A trained conditional probability model $P(tag|suffix)$ (e.g. $P(VBG|ing)$, $P(N|ing)$, etc.) was used to predict the unknown words based on a words' ending information. The model was trained based on a pair of "word-tag" frequencies in the dictionary; if the word had an ending that fell in the top 100-words' ending list.

In order to estimate the initial emissions for unknown words, a paradigmatic similarity measure was used which estimated the conditional probabilities of the POS tag when given an unknown word; i.e. $P(w|t)$ (Cucerzan *et al.*, 2000). The estimation depended on the words' ending information to characterise similar words (words with similar features). The other method to estimate the initial emissions for unknown words would be an artificial annotated corpus created to induce "word-tag" frequency by using the label propagation graph (Garrette *et al.*, 2012). The automatic generation of corpus annotation was able to reduce the noise in the artificial dictionary (expended tagged dictionary). Furthermore, two potential sources of knowledge (i.e. tagged dictionary and raw text sequences) were used to estimate the initial emission of unknown words. $P(w|t)$, which was an early intuitive initial emission, provided the basic need to run expectation maximisation. The label propagation graph was subsequently modified by adding or replacing its generic words' ending information with a focused set of the words' ending features generated by Finite State Transducers (Garrette *et al.*, 2013).

In order to estimate the lexical probabilities for ambiguous words, the lexicon probabilities $P(t|w)$ were assigned by either a linear-context-based or an ambiguity-class guesser (Goldberg *et al.*, 2008). The ambiguity-class guesser assigned each unknown word with a set of open-class tags that appeared with the words' ending in the dictionary. The words' ending was set to the longest setting of up to three characters and limited to the top 100-words' ending in the dictionary. Exploiting the words' ending information by treating it as a feature to a model was also applied in other approaches. For example, the prototype-driven learning approach (Haghighi *et al.*, 2006), Conditional Random Fields learning approach (Subramanya *et al.*, 2010), cross-lingual POS tagging approach (Das *et al.*, 2011; Täckström *et al.*, 2013) and Bayesian approach based on Latent Dirichlet allocation (Toutanova *et al.* , 2007; Hasan *et al.*, 2009).

Therefore, this dictionary was attributed to an artificial one since it was a filtered dictionary where all words, with only tags that had certain relative frequencies, occurred in the annotated corpus that were taken as entries. This provided a scenario where every word in the corpus was considered as known words. In contrast to this scenario, for under-resourced languages such as Malay, a tagged corpus is not freely available. Therefore, a similar dictionary was unable to form; yet, some printed dictionaries were available and used as an alternative. However, the printed dictionary does not include most of the words that exist in the corpus, especially passive verbs and derivative words. This caused such words to be considered as unknown since the words that even appear in the corpus have unknown tags.

2.0 RELATED WORK

Several researchers had conducted their efforts to train unsupervised HMM POS tagger, which catered words not listed in the dictionary or ambiguous words. They exploited the words' ending with specified length to enlarge the training dictionaries (Miller *et al*, 2007; Ravi *et al.*, 2009); or directly estimated the initial emissions for unknown words (Cucerzan *et al.*, 2000; Garrette *et al.*, 2012); or directly estimated the lexical probabilities for ambiguous words (Goldberg *et al.*, 2008). There are also researchers who automatically built an annotated corpus and then trained the HMM using the supervised approach (Garrette *et al.*, 2013).

In order to enlarge the training dictionary, completely unknown lexicon w was created by exploiting the words' ending patterns as the exemplary, which was extracted from an example corpus, such as the WSJ corpus (Miller *et al.*, 2007). For each unknown lexical w, all immediate neighbours (words similar to w from the annotated corpus) were searched. The average of the exemplary lexical probability was used to obtain the lexical probability of w; i.e. $P(t|w)$. This method was very effective; even the application domain was significantly different from the exemplary text. The other method was to enlarge the training dictionary. The unknown words with their predicted POS were added as entries (Ravi *et al.*, 2009). A trained conditional probability model $P(tag|suffix)$ (e.g. $P(VBG|ing)$, $P(N|ing)$, etc.) was used to predict the unknown words based on a words' ending information. The model was trained based on a pair of "word-tag" frequencies in the dictionary; if the word had an ending that fell in the top 100-words' ending list.

In order to estimate the initial emissions for unknown words, a paradigmatic similarity measure was used which estimated the conditional probabilities of the POS tag when given an unknown word; i.e. $P(w|t)$ (Cucerzan *et al.*, 2000). The estimation depended on the words' ending information to characterise similar words (words with similar features). The other method to estimate the initial emissions for unknown words would be an artificial annotated corpus created to induce "word-tag" frequency by using the label propagation graph (Garrette *et al.*, 2012). The automatic generation of corpus annotation was able to reduce the noise in the artificial dictionary (expended tagged dictionary). Furthermore, two potential sources of knowledge (i.e. tagged dictionary and raw text sequences) were used to estimate the initial emission of unknown words. $P(w|t)$, which was an early intuitive initial emission, provided the basic need to run expectation maximisation. The label propagation graph was subsequently modified by adding or replacing its generic words' ending information with a focused set of the words' ending features generated by Finite State Transducers (Garrette *et al.*, 2013).

In order to estimate the lexical probabilities for ambiguous words, the lexicon probabilities $P(t|w)$ were assigned by either a linear-context-based or an ambiguity-class guesser (Goldberg *et al.*, 2008). The ambiguity-class guesser assigned each unknown word with a set of open-class tags that appeared with the words' ending in the dictionary. The words' ending was set to the longest setting of up to three characters and limited to the top 100-words' ending in the dictionary. Exploiting the words' ending information by treating it as a feature to a model was also applied in other approaches. For example, the prototype-driven learning approach (Haghighi *et al.*, 2006), Conditional Random Fields learning approach (Subramanya *et al.*, 2010), cross-lingual POS tagging approach (Das *et al.*, 2011; Täckström *et al.*, 2013) and Bayesian approach based on Latent Dirichlet allocation (Toutanova *et al.* , 2007; Hasan *et al.*, 2009).

It was noted that expanding a Malay tagged dictionary using any tagged corpus (Miller *et al.*, 2007) would not be easily implemented because a similar Malay tagged corpus was not freely available. Even if some Malay text was tagged to induce a pair of "tag-words' ending", used to assign the initial value of emission probability of HMM for an unknown word, it was doubtful in regards to the accuracy of the trained transition probabilities of the HMM later. It was proven that assigning the initial value to the transition probabilities of tag sequences for unambiguous words would increase the tagging accuracy (Banko *et al.*, 2004). These trained transitions were more reliable as compared to the transitions for unknown word tags. The similar solution to induce "tag-words' ending" dictionary (Ravi *et al.*, 2009) also raised doubts since the distribution of a words' ending in the dictionary did not represent their distribution in the real corpus. For example, the circumfix of Malay *di-...-kan* is very rare in the dictionary (excluded from the top 100-affix list) but repeated more than a thousand in the real corpus, which was affixed to unknown words.

It was viewed that in contrast to the distribution of "word-tag" in the dictionary to induce "tag-words' ending" frequencies (Ravi *et al.*, 2009), it would be worthwhile to take into account the expected distribution of "word-tag" from the untagged corpus. In supervised HMM tagger, "word-tag" frequencies came from the annotated corpus (Brants, 2000). Therefore, for our unsupervised HMM Malay tagger, which used the untagged corpus, the distribution of "word-tag" required estimation. For emission initialisation purposes, all words in the training corpus found in the dictionary were mapped with possible tags. On the other hand, words not found in the dictionary were mapped with 'unknown' tags ($t_{unknown}$). It was assumed that an unknown tag ($t_{unknown}$) was one of the tags in the HMM in which the emission probabilities $P(w|t_{unknown})$ were also produced in conjunction with other emissions during the training. Furthermore, the substitution method was used to replace them with some value produced by way of handling unknown words mechanism.

Malay is an agglutinative language, which has rich morphology to derive other words (Nik Safiah *et al.*, 2010; Abdullah, 2006). Affixation is most commonly used for the morphological process, and three types of affixations are present; i.e. prefixes, suffixes and circumfixes. These affixations can always be intuitively used for guessing the POS of derived words. Therefore, research in the field of part of speech tagging in Malay must also emphasise the morphological characteristics of the Malay origin as opposed to the traditional basic statistical POS tagging, which is linguistically independent and does not explicitly include linguistic features. This study aimed to examine the effectiveness of using actual affix morphemes of the Malay language as compared to the use of the words' ending characters as features for predicting the POS of unknown words. Therefore, a suitable combination of methods in the unsupervised HMM tagging for Malay was recommended.

3.0 MALAY TAG SET

The ideal number of tags for statistical POS tagger must be in a small number to reduce ambiguousness and increase accuracies (Brants, 1995; Dienes *et al.*, 2000; Dominguez *et al.*, 2008; Petrov *et al.*, 2012). The number of tags can be reduced by combining some sub-classes under a main class into one tag (Merialdo, 1994; Azimizadeh *et al.*, 2008; Mohseni *et al.*, 2008). Even though the more the tag number encoded more linguistic information, such as morphological and syntactic structures of words, it created difficulties to distinguish similar POS tags (Güngör, 2010). Tiny differences to the tags of the same sub-class, which is too detailed, would result in inconsistencies in tagging and would have difficulties to acquire consensus among annotators. Annotations performed to Penn Treebank WSJ text indicated that 7.2% of the cases involved disagreement among annotators (Marcus *et al.*, 1993). Thus, too many sub-classes cause inconsistencies in the tag set which impairs the ability of the taggers (Güngör, 2010). Therefore, in this study, the number of Malay tags was approximated to the number of the Penn Treebank tags in which the number was assumed as suitable, and it was tested with the existing standard of HMM with expected performance.

In order to define a linguistically motivated Malay tag set, while at the same time determining a suitable number of tags for the statistical approach, the classification of Malay words described in Malay grammar textbooks written by many scholars were referred (Nik Safiah, 2010; Abdullah, 2006; Abdullah *et al.*, 2006;

Asmah, 2009). The descriptions provided guidance in designing a tag set, rather than trying to adopt from English or other languages. Accordingly, many monolingual or bilingual Malay dictionaries associated word classes of entries, which tallied with grammar descriptions in those books (Hock, 2009; Hawkins, 2008; Arbak, 2005). Therefore, the word class descriptions in the textbooks were used to refine word classes used in the dictionary of (Arbak, 2005). The modifications on the tag sets in the dictionary were as follows:

(a) Designated the auxiliary (*Kata Bantu*) into two classes: modal auxiliary (*Kata Bantu Ragam*) and aspectual auxiliary (*Kata Bantu Aspek*).

The modal auxiliary described the mood of the acts on the verbs; for example, *hendak* (want), *mahu* (wish), *harus* (should), *mesti* (must), *boleh* (can) and *dapat* (can). There are no clear verb tenses in Malay as opposed to English, therefore, aspectual auxiliaries differentiate a verb, whether it had already past, still on-going or yet to be done; for example, *telah* (already past), *sudah* (already past), *pernah* (ever), *sedang* (still), *masih* (still), *akan* (will) and *belum* (not yet) were used.

(b) Fine-grained function words (*Kata Tugas*) according to their roles.

Function words in Malay are limited but they significantly play different roles in a text. They are used in a sentence or phrase as a grammatical function. Their role can be as determiners (*Kata Penentu*), imperative words (*Kata Perintah*), discourse markers (*Penanda Wacana*), affirmative words (*Kata Pembenar*), directional words (*Kata Arah*), assertion words (*Kata Penekan*) and nominalisers (*Kata Pembenda*).

(c) Created existential (*Kewujudan*) word class.

This tag was created to differentiate between verbs in dominantly *Subject-Verb-Object* Malay sentence patterns with another special case verb, *ada* (exist), in the sentence pattern of *Verb-Subject* which is very rare in Malay. For example, the sentence '*ada lima puluh ekor kambing*' (there are fifty goats) complies with the Verb-Subject pattern as compared to '*dia ada lima puluh ekor kambing*' (he/she has fifty goats), which complies with the *Subject-Verb-Object* pattern.

(d) Allocated relative pronouns (*Ganti nama relatif*) for *yang* (which, that).

Many Malay grammar textbooks classify the word *yang* as a relative subordinating conjunction (*kata hubung pancangan relatif*), which is the subclass of *kata hubung* (conjunction). According to our corpus, the word *yang* was the most frequently used. So in statistical POS tagger, it was better to classify such words in one class in such a way that it does not affect the transition probability of HMM caused by highly used words in a class, as opposed to rarely used words of the same class.

There were forty (40) Malay tag-sets used in our tagger, including symbol and punctuation tags, which resulted after conducting the above modification. This number was comparable to the number of tags in the Penn Treebank tag set, which was considered as suitable for HMM and the performance was expected. A complete set of Malay tags were listed in Table 1.

4.0 TAGGING UNKNOWN WORDS

Unknown words often play an important role in describing the meaning of a sentence than known words because an unknown word is mostly a special word that carries more semantic information than a known word (Vadas *et al.*, 2005). Most of the unknown words were assigned with an open class, such as nouns or verbs, by the assumption that they are impossible to exist in close class category such as determiners or prepositions. Handling unknown words was considered as key to improve the performance of POS taggers (Hall, 2003). POS tagging models that are able to handle unknown words are often used and adapted into tagging any under-resourced languages (Dandapat, 2009).

Table 1: Malay Tag Set

No	Tags	Description	Examples
1	ADA	Existential (*Kewujudan*)	*ada* (exist)
2	GEL	Title (*Gelaran*)	*Datuk, Haji*
3	GNR	Relative pronoun (*Ganti nama relatif*)	*yang* (which, that)
4	JDH	Numeral classifier (*Penjodoh Bilangan*)	*orang* (people), *buah* (fruit)
5	KA	Adjective (*Kata Adjektif*)	*pandai* (clever), *bodoh* (stupid)
6	KAD	Adverb (*Kata Adverba*)	*sekarang* (now), *tadi* (justnow)
7	KAR	Directional Word (*Kata Arah*)	*bawah* (under), *tepi* (side)
8	KBA	Aspectual Auxiliary (*Kata Bantu Aspek*)	*akan* (will), *belum* (notyet)
9	KBIL	Numeral (*Kata Bilangan*)	*satu* (one), *100*
10	KBR	Modal Auxiliary (*Kata Bantu Ragam*)	*mahu* (wish), *harus* (should)
11	KEP	Abbreviation (*Kependekan*)	*UKM*
12	KGN	Pronoun (*Kata Ganti Nama*)	*kamu* (you), *awak* (you)
13	KH	Conjunction (*Kata Hubung*)	*dan* (and), *lalu* (then)
14	KK	Verb (*Kata Kerja*)	*pulang* (return), *tidur* (sleep)
15	KN	Common Noun (*Kata Nama Am*)	*rumah* (house), *kambing* (goat)
16	KNF	Negative Word (*Kata Nafi*)	*bukan* (not) *and tidak* (no/not)
17	KNK	Proper Noun (*Kata Nama Khas*)	New York, Pasir Mas
18	KP	Intensifier (*Kata Penguat*)	*Sungguh* (true/exact)
19	KPB	Nominalizer (*Kata Pembenda*)	*lajunya* (its speed), *sakitnya* (painfulness)
20	KPM	Narrator (*Kata Pemeri*)	*ialah* (is), *adalah* (is)
21	KPN	Emphatic Words (*Kata Penegas*)	*juga* (also), *jua, pun*
22	KPR	Affirmative Word (*Kata Pembenar*)	*ya* (yes), *benar* (true)
23	KPT	Assertion Word (*Kata Penekan*)	*nampaknya* (it seems), *bahawasanya*
24	KS	Preposition (*Kata Sendi*)	*dari* (from), *pada* (at)
25	KSR	Interjection (*Kata Seru*)	*amboi* (wow), *bedebah* (ah)
26	KTP	Imperative Word (*Kata Perintah*)	*sila* (please), *jemput* (invite)
27	KTY	Interrogative Word (*Kata Tanya*)	*berapa* (how), *bila* (when)
28	PIN	Foreign Word (*Perkataan Asing*)	*university*
29	PW	Discourse Marker (*Penanda Wacana*)	*kalakian* (urging), *maka* (then)
30	TEN	Determiner (*Kata Penentu*)	*ini* (this) *and itu* (that)
31	EMAIL	Email Address/ Web Site (*Alamat email / alamat halaman sesawang*)	hassan.dbangi@yahoo.com
32	$	Dollar Sign (*Tanda Dollar*)	$ RM
33	#	Pound Sign (*Tanda Pound*)	# £
34	“	Left Quote (*Tanda Pembuka Kata*)	‘ “
35	(	Left Parenthesis (*Tanda Buka Kurungan*)	([{ <
36	)	Right Parenthesis (*Tanda Tutup Kurungan*)	)] } >
37	,	Comma (*Tanda koma*)	,
38	.	Sentence-Final Punctuation (*Tanda Penutup Ayat*)	! ? .
39	:	Mid-Sentence Punctuation (*Tanda Pertengahan Ayat*)	- ... ; :
40	SYM	Any Symbols (*Sebarang Simbol*)	` ^ _ @ * / \ & % + = | ~

The term "unknown word" in statistical POS tagging refers to words that are not in the training corpus or dictionary. Therefore, for unsupervised HMM POS tagging, words that are not in the training corpus are known as unseen words. The unseen words lead to the problem of inexistence of their emission probabilities, which requires the Viterbi tagging to approximate the value with some mechanism. However, words that are not found in the dictionary will lead to a case where the words that appear in the corpus cannot be matched to their tags, in which their initial emission probabilities must also be approximated with some mechanism. In this case, the substitution method was used to replace emission probability of words not listed in the dictionary, i.e. $P(w|t_{unknown})$, with some value produced by the mechanisms of handling unknown words; this was discussed in section 4.1, 4.2 and 4.3, after the tag $t_{unknown}$ was assumed as one of the tags in HMM during training.

In order to assign the initial emissions of HMM training, the words were grouped into their equivalent classes after being matched with POS tags referred in the dictionary. For example, words that were categorised as only nouns were pooled into one equivalent class, and words that can be categorised as either verbs or adjectives were in another class, and so on. Grouping the words into their equivalent classes tremendously reduced the number of emission parameters, giving an advantage in estimating the transitions more reliably (Banko *et al.*, 2004; Kupiec, 1992). Any word that occurred in the corpus more than 100 times was individually grouped into a single class to reduce the skewed probability of rare words by high frequency words within the same class. Each group was treated as a metaword, u_L, where L was a subset of the integers from 1 to T, and T was the number representing the tags.

The HMM training employs an untagged Malay corpus containing 995,240 tokens, and consisting of 30,640 word types; including symbols. Out of 30,640 word types, 14,068 words were not listed in the dictionary. After completing the training, the metaword grouped under the 'unknown' tag also had an emission probability, which was the probability of the metaword given by an 'unknown' tag, $P(u_L|t_{unknown})$. The value of these probabilities was produced in conjunction with other emission values by the forward-backward training. Nevertheless, these values were substituted with certain probabilistic measures, discussed in section 4.1, 4.2 and 4.3.

For every iteration in the forward-backward training, the trained emission probability $P(u_L|t_i)$ was proportionate to the probability of the word given a tag $P(w|t_i)$ by the assumption $(w|t_i) = \frac{C(w)}{C(u_L)} P(u_L|t_i)$ if $w \in u_L$, where $C(w)$ was the number of token w and $C(u_L)$ was the total token accumulated in metaword u_L. Hence, the joint probability of metawords u_L with tag t_i was estimated as follows:

$$P(u_L, t_i) = P(u_l|t_i)P(t_i) \tag{1}$$

The marginal $P(t_i)$ was estimated using the following equation:

$$\begin{aligned} P(t_i) &= \frac{expected\ number\ of\ times\ in\ state\ i}{expected\ number\ of\ times\ in\ all\ states} \\ &= \frac{\sum_{o=1}^{O} \gamma_o(t_i)}{\sum_{j=1}^{T} \sum_{o=1}^{O} \gamma_o(t_j)} \end{aligned} \tag{2}$$

$\gamma_o(t_i)$ was the probability of being in state t_i at observation o for a given observation sequence in the HMM model. The probability of a metaword $P(u_L)$ was calculated after grouping the words into metawords and dividing the number of a metaword over all metawords:

A reverse conditional probability $P(t_i|u_L)$ was as follows:

$$P(t_i|u_L) = \frac{P(u_L, t_i)}{P(u_L)} \tag{3}$$

The number of joint "word-tag" was estimated as follows:

$$C(w_k, t_i) = P(t_i|w_k)C(w_k) \tag{4}$$

The $P(t_i|w_k)$ in (4) was substituted by $P(t_i|u_L)$ for every $w_k \in u_L$. Therefore:

$$C(w_k, t_i) = \frac{P(t_i|u_L)C(w_k)^2}{C(u_L)}; \quad \forall w_k \in u_L \tag{5}$$

The number of word type $C(w_k)$ was counted based on the word type's frequency in the training corpus.

4.1 Morpheme-based POS Guessing

The way of forming derivative words in Malay is accomplished by merging root words with affixes. For example, a root word *serap* (absorb), can produce new words such as *menyerap* (absorb), *menyerapkan* (induct), *diserapkan* (inducted), *menyerapi* (permeated), *diserapi* (be permeated), *penyerap* (absorber), *penyerapan* (absorption), *terserap* (absorbed), *terserapkan* (absorbable), *serapan* (absorption), *keterserapan* (absorptive), *daya serap* (absorptive) and kedayaserapan (absorptiveness). Affixes are considered bound morpheme as opposed to root words, which are unbound morpheme that can receive affixations. Therefore, affixes cannot be present alone in a sentence (for example, *ber-*, *ter-*, *ke-*, *me-*, *-nya*, *-kah*, *-lah*, *-pun*, *-an*), they must be affixed to root words. Affixes can be categorised into three types, i.e. prefixes, suffixes and circumfixes.

Prefixes can exist in derivative nouns such as *pembuat* (manufacturer), derivative verbs such as *mendaki* (climb) and derivative adjectives such as *terendah* (lowest). Suffixes can exist in derivative nouns such as *ukuran* (measurement) and derivative verbs such as *besarkan* (enlarge). Circumfixes can exist in derivative nouns such as *pembinaan* (construction), derivative verbs such as *mendermakan* (donate) and derivative adjectives such as *kecinaan* (*Chineseness*). The part of speech (POS) of many derivative words formed by Malay morphological rules were predicted in a way in which derivative nouns as *Kata Nama* (Noun) were represented as KN, derivative verbs as *Kata Kerja* (Verb) were represented as KK and derivative adjective as *Kata Adjektif* (Adjective) were represented as KA. The morphological rules were outlined as follows:

Rule 1:

POS = {'KN'} if the derivative word has any following affixes:

1. Circumfixes: { *per-...-an*, *penge-...-an*, *peng-...-an*, *pen-...-an*, *pem-...-an*, *pel-...-an*, *pe-...-an* }
2. Prefixes: { *tata-...*, *supra-...*, *sub-...*, *pra-...*, *per-...*, *penge-...*, *peng-...*, *pen-...*, *pem-...*, *pel-...*, *pe-...*, *maha-...*, *ke-...*, *juru-...*, *eka-...*, *dwi-...* }
3. Suffixes: { *...-wati*, *...-wan*, *...-man*, *...-isme*, *...-in*, *...-at*, *...-an*, *...-ah* }

Rule 2:

POS = {'KK'} if the derivative word has any following affixes:

1. Circumfixes: { *menge-...-kan*, *meng-...-kan*, *meng-...-i*, *men-...-kan*, *men-...-i*, *memper-...-kan*, *memper-...-i*, *mem-...-kan*, *mem-...-i*, *me-...-kan*, *me-...-i*, *ke-...-an*, *diper-...-kan*, *diper-...-i*, *di-...-kan*, *di-...-i*, *ber-...-kan*, *ber-...-an* }
2. Prefixes: { *meny-...*, *menge-...*, *meng-...*, *men-...*, *memper-...*, *mem-...*, *me-...*, *diper-...*, *di-...*, *ber-...*, *bel-...*, *be-...* }
3. Suffixes: { *...-kan*, *...-i* }

Rule 3:

POS = {'KA'} if the derivative word has any following prefixes:

1. Prefixes: { *te-...*, *se-...* }

Rule 4:

POS = {'KN', 'KA'} if the derivative word has the following circumfix:

1. Circumfix: { *ke-...-an* }

Rule 5:

POS = {'KK', 'KA'} if the derivative word has the following prefix:

1. Prefix: { *ter-...* }

Applying the linguistic rules above is quite crucial in terms of deciding the precedence of affixes for the best guessing of word classes. However, if we examine the letters in the affixes, the longest affix string can become a superset to the shorter one. For example, the prefix *pe-...* in *Rule 1* is a superset to the prefix *per-...*, *penge-...*, *peng-...*, *pen-...*, *pem-...* and *pel-...* in terms of characters in the strings. Therefore, the longest affixes are always put at the highest precedence. To accomplish this idea, each set of circumfixes, prefixes and suffixes in the rules were sorted into their descending orders. The circumfixes were made up of both certain prefixes and suffixes, in which both prefixes and suffixes were subsets to circumfixes. For example, the circumfix *diper-...-kan* is made up of a combination of the prefix *diper-...* and suffix *...-kan*. Therefore, the circumfix became the highest precedence followed by prefixes and suffixes for guessing the class of unknown words. The suffixes became lower precedence compared to prefixes because there were fewer suffixes than prefixes.

To facilitate the integration of morphological rules into HMM Malay POS tagger, the directed graphs were used to represent the rules. Therefore, Fig. 1 represented the circumfixes of Rule 1, 2 and 4; Fig. 2 represented the prefixes of Rule 1, 2, 3 and 5; and Fig. 3 represented the suffixes of Rule 1 and 2. The red nodes indicated the start of tracking prefixes in Fig. 2, whereas the blue nodes indicated the start of tracking suffixes in Fig. 3. Circumfixes in Fig. 1 are successfully tracked if tracking both prefixes and suffixes met at the determinant nodes indicated by the orange colour in which the predicted POS was embedded. Similarly, prefixes and suffixes are successfully tracked if the tracking met at determinant node (orange colour). An algorithm to guess the POS of unknown words using morphological rules was given in Fig. 4. This algorithm was used to examine the existence of Malay affix morphemes in unknown words and then predict their POS. Furthermore, the algorithm was treated as a baseline tagging in this study. The baseline tagging does not involve any training corpus or tagged dictionary; however, all tokens in the test corpus were tagged based-on the affixation information.

Since the baseline tagging was not hybridised with other tagging methods, the algorithm depended only on characters in the affix morphemes, therefore, there was no mechanism to disambiguate the POS of words having more than one tag. To resolve this ambiguousness, the POS of such words were assigned with the first element of POS set returned by the algorithm, which in most cases, were *Kata Nama* (noun) or *Kata Kerja* (verb).

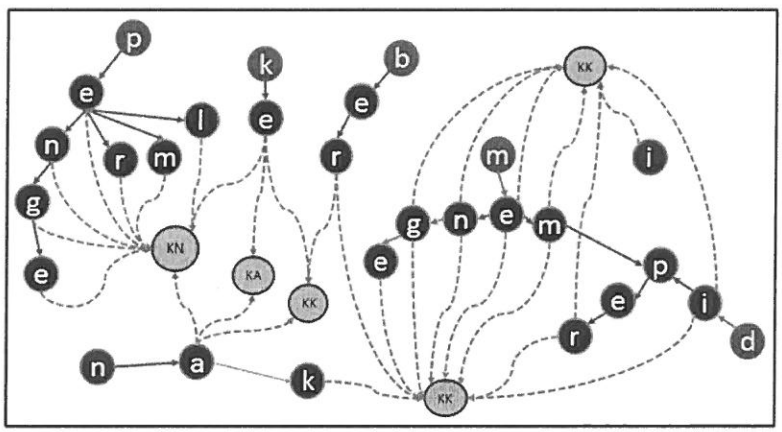

Fig. 1 Circumfix graph

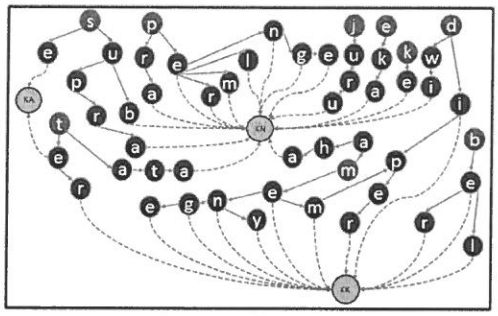

Fig. 2 Prefix graph

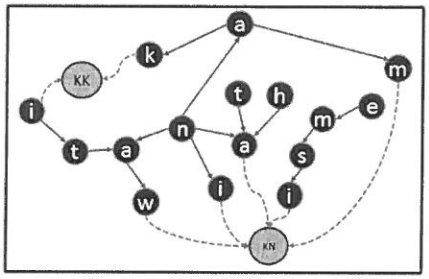

Fig 3. Suffix graph

For each unknown words, find their affix morpheme using the following steps:

1. Travers the circumfix graph
   If meet determinant node then
   Return POS set embedded to the node
2. Else travers prefix graph
   If meet determinant node then
   Return POS set embedded to the node
3. Else travers suffix graph
   If meet determinant node then
   Return POS set embedded to the node
4. Else
   Return POS set = { 'KN', 'KNK', 'KK' }

Fig 4. POS guesser algorithm using affix morphemes

The POS guesser algorithm was integrated with unsupervised HMM for tagging Malay unknown words. Whenever the tagger came cross unknown words, the tags for those words were allocated with possible tags given by the POS guesser algorithm. However, HMM tagging needs words' emission probability to disambiguate and assign the most possible POS tags according to the context of words. Since unknown words are not seen in the training corpus, such emission values were missing. Therefore, to resolve this issue, the emission probabilities were estimated in two ways. First, the emission probabilities of unknown words were assigned according to uniform distribution of all possible tags given in (6), where X was a set of possible POS of the unknown word returned by the POS guesser algorithm, $|T|$ was the number of all tags ($|T| = 40$) and δ was a smoothing factor in which the best value was 0.01. This value was observed through the experimental result using the development corpus (30,017 tagged-tokens). The experiment was conducted in a cross validation observation in which the development corpus was partitioned into ten partitions with similar size (about 3K each). Nine of them were merged back and used for training the model and the rest were used for observation. This process was repeated ten times, such that each partition was used for training and observation. Table 2 showed the different values given to δ against the accuracies of tagging the unknown words in each partition. According to observations, the given value 0.01 to δ showed

the best, whereby, any merging partitions of the development corpus always gave high accuracy of tagging unknown words in observed partitions.

Second, the emission probabilities were assigned according to marginal proportionate distribution of tags produced during HMM training given in (7), where $P(t)$ was the probability of tag; Y was the normalisation factor; and δ was the smoothing factor defined as the lowest $P(t)$ for t in X multiplied by coefficient ϵ ($\epsilon = 0.1$ was the best value observed). This observed value was also determined by a cross validation observation. Table 3 presented the different values given to ϵ against the accuracies of tagging the unknown words in each partition. According to observations, the given value of ϵ as 0.1 displayed the best, whereby, any merging partitions of development corpus always yielded high accuracy of tagging unknown words in observed partitions.

$$P(w|t) \cong \begin{cases} \dfrac{1+\delta}{|X|+\delta|T|} & if\ t \in X \\ \dfrac{\delta}{|X|+\delta|T|} & if\ t \notin X \end{cases} \tag{6}$$

$$P(w|t) \cong \begin{cases} \dfrac{P(t)+\delta}{Y}, & if\ t \in X \\ \dfrac{\delta}{Y}, & if\ t \notin X \end{cases} \tag{7}$$

Table 2 Observation results for tagging unknown words in each partition against different given δ values

Observing corpus	**Given δ values**				
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Partition 1	32.14%	38.74%	37.62%	37.61%	37.61%
Partition 2	31.22%	37.74%	36.01%	35.32%	35.10%
Partition 3	32.32%	38.05%	37.01%	36.82%	36.70%
Partition 4	33.00%	38.73%	37.62%	37.61%	37.61%
Partition 5	31.82%	38.64%	38.00%	37.80%	37.80%
Partition 6	32.00%	37.84%	36.61%	35.82%	35.80%
Partition 7	31.00%	37.54%	36.91%	35.72%	35.50%
Partition 8	31.23%	37.94%	36.91%	36.32%	36.10%
Partition 9	32.24%	38.77%	37.52%	37.51%	37.51%
Partition 10	32.10%	38.70%	37.82%	37.31%	37.11%

Table 3 Observation results for tagging unknown words in each partition against different given ϵ values

Observing corpus	**Given ϵ values**				
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Partition 1	39.31%	38.26%	37.86%	37.82%	37.80%
Partition 2	38.50%	37.87%	37.18%	37.11%	37.08%
Partition 3	39.51%	38.37%	37.97%	37.90%	37.85%
Partition 4	39.61%	38.36%	37.96%	37.83%	37.80%
Partition 5	38.90%	38.01%	37.52%	37.08%	37.00%
Partition 6	38.90%	38.00%	37.55%	37.49%	37.45%
Partition 7	38.40%	37.87%	37.17%	37.10%	37.08%
Partition 8	38.80%	37.81%	37.56%	37.49%	37.40%
Partition 9	39.50%	38.10%	37.25%	37.20%	37.19%
Partition 10	39.00%	38.00%	37.20%	37.19%	37.15%

4.2 Predicting POS through a Words' Starting

The term "words' starting" is the sequence of characters that begin a word string. For example, the word "*hasut*" (instigate) can have a word starting set of {"*h*", "*ha*", "*has*", "*hasu*"} for a predefined length of four characters. Intuitively, the longer the sequence of characters, the stronger the judgment in predicting a particular words' tag. For substantial amounts of this information, they are able to induce alternative emission probability values of unknown words. For example, the probability of a tag given a words' starting was estimated based on the statistical data available for words that begin with the same sequence of letters. Therefore, the probability distribution can be generated from all words in the training corpus that share the same sequence of letters for some predefined length. This model implicitly embedded the linguistic knowledge of Malay affixes. The probability of a tag t_i given the first m letters $l_1 l_2 \dots l_m$ of the letter sequence in a word was estimated and smoothed using successive abstraction (Brants, 2000; Samuelsson, 1996). This estimation was recursively calculated by considering the marginal distribution of tags $P(t_i)$ produced by forward-backward training, formulated in (2), and the standard division in (14) to every successive character.

$$\hat{P}(t_i|l_1 l_2 \dots l_m) = \frac{P(t_i|l_1 l_2 \dots l_m) + \sigma \hat{P}(t_i|l_1 l_2 \dots l_{m-1})}{1 + \sigma} \tag{8}$$

$$\hat{P}(t_i|l_1) = \frac{P(t_i|l_1) + \sigma \hat{P}(t_i)}{1 + \sigma} \tag{9}$$

$$P(t_i|l_1\, l_2 \dots l_m) = \frac{C(t_i,\ l_1 l_2 \dots l_m)}{C(l_1 l_2 \dots l_m)} \tag{10}$$

$$\hat{P}(t_i) = P(t_i) \tag{11}$$

For any defined length of m and $m > 0$, $C(t_i,\ l_1 l_2 \dots l_m)$ was the total number of joint "word-tag" that shared the same words' starting $l_1 l_2 \dots l_m$ with tag t_i; $C(l_1 l_2 \dots l_m)$ was the total number of word types that shared the same words' starting $l_1 l_2 \dots l_m$. Therefore:

$$C(t_i, l_1 l_2 \dots l_m) = \sum_{w \subset l_1 l_2 \dots l_m} C(w, t_i) \tag{12}$$

$$C(l_1 l_2 \dots l_m) = \sum_{w \subset l_1 l_2 \dots l_m} C(w) \tag{13}$$

The number of word type $C(w)$ was easily counted based on the word type's frequency in the training corpus. However, the number of joint "word-tag" $C(w, t_i)$ was estimated using (5) and the value of σ was a standard deviation of the marginal distribution of tags (14) produced in each iteration of forward-backward training.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(P(t_i) - \bar{P})^2} \tag{14}$$

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P(t_i)$$

4.3 Predicting POS through a Words' Ending

A similar concept to section 4.2 was used for predicting the POS of unknown words through a words' ending by looking at the backward sequence of characters that make up a word. For example, the word "*hasut*" can have a word ending set of {"*t*", "*ut*", "*sut*", "*asut*"} for a predefined length of four characters. The probability distribution for a particular words' ending was generated from all word types in the training corpus that shared the same words' ending of some predefined length. The probability of tag t_i given the last m letters $l_{n-m+1}l_{n-m+2} \dots l_n$ of a word was recursively estimated and smoothed, similar to predicting POS for unknown words through a words' starting in section 4.2, but treated in reverse order of characters.

5.0 EXPERIMENTAL RESULTS

The accuracy of the tagging denoted the percentage of the words correctly assigned with tags as compared to the tagged corpus (Schröder, 2002). Therefore, the tagging performance was often measured by the overall tagging, known word and unknown word tagging accuracies (Dandapat, 2009; Giesbrecht *et al.*, 2009). Known words always referred to words seen in the training corpus and unknown words referred to words not seen in the corpus. However, in this case, the definition of unknown words was extended to include the words that may exist in the training corpus but not listed in the dictionary. Therefore, the accuracy in our evaluation was designates into five types of accuracies to ease the analysis of tagging:

(a) Overall – the overall performance of the tagger.
(b) Seen word with unique tag – the performance of tagging words seen in the training that exist in the dictionary with only one tag.
(c) Seen words with ambiguous tags– the performance of tagging words seen in the training that exist in the dictionary with more than one tag.
(d) Seen words not existing in the dictionary – the performance of tagging words not listed in the dictionary but seen in the training.
(e) Unseen words – the performance of tagging words unseen in the training corpus.

Each accuracy was calculated as the ratio of correctly tagged words of the related accuracy in the test corpus to the total number of all tagged words of the related accuracy in the test corpus. Table 4 presented the results of the experiments.

5.1 HMM-Viterbi Tagging with a Words' Starting or Words' Ending

This method of tagging unknown words was considered as a character-based prediction as opposed to morphological rules. There were two possibilities that influenced the results, i.e. the number of training iterations and the maximum predefined length of characters used in the words' starting or ending. During the test, one of the other must be a constant. Due to that reason, the experiment was repeated for each predefined length of characters (ranging from one to twelve characters) for the words' starting and ending methods, with various numbers of iterations (ranging from one to ten iterations). The iteration and the length of characters that gave the highest overall performance were considered as the best performance. Therefore, the experiments showed that the best performance was with a maximum predefined length of four (4) characters for words' starting method on the second iteration of HMM training. The overall accuracy was 81.81%. Similarly, the maximum predefined length of words' ending method with eight (8) characters on the third iteration of HMM training had an overall accuracy of 81.71%. Although the percentage did not show significant difference, there were differences in terms of the number of tokens with 0.1% of accuracy, reflecting about 121 tokens (out of 121,090 test tokens), which provided a significant comparative number. The performance trends of both, using the words' starting method or words' ending method, were shown in Fig. 5.

A comparison of the performance of tagging unseen words using the words' starting method was shown in Fig. 6. The tagging accuracy for unseen words by using a words' starting information was 39.42% on the fourth iteration of HMM training. On the other hand, using a words' ending information, the accuracy was 33.22% on the second iteration. The percentage showed a difference of 7.20%, indicating that using a

words' starting information was slightly more accurate than using a words' ending information. Furthermore, using a words' ending information required more characters.

Tagging seen words not listed in the dictionary using a words' starting information outperformed the use of a words' ending information, as shown in Fig. 7. The tagging accuracy for using a words' starting information was 39.02% on the third iteration, as compared to using a words' ending information, which was 38.36% on the fourth iteration. The difference of 0.66% reflected about 105 tokens (out of 15,882). This finding strengthened the argument to use a words' starting information for character-based prediction of unknown words' POS.

5.2 HMM-Viterbi Tagging with Unknown Words' POS Predicted through Malay Affix Morphemes

The number of training iterations can influence results, so the experiments were repeated for each iteration ranging from one to ten. The best overall performance from those iterations was considered the best result. Table 4 showed the results of tagging performance using a combination of HMM-Viterbi with handling unknown words using morpheme-based POS guessing (stated in row 4 and 5). The best overall tagging accuracy was 82.28% when the emission was substituted by a value proportionate to the marginal distribution of tags. The results were slightly better than the results of the experiment done on HMM-Viterbi with words' starting or ending methods. However, tagging unseen words was less accurate in HMM-Viterbi tagging with morpheme-based POS guessing (31.94%) compared to the results of the experiment done on HMM-Viterbi with words' starting method (33.42%), with a difference of about 1.48% (see row 2 and 5 of Table 4). This indicated that the affixation rule did not enhance the accuracy of POS guessing of unseen words. On the other hand, tagging words identified as not in the dictionary was slightly more accurate on Viterbi tagging with morpheme-based POS guessing, with an accuracy of 42.52%; better than the baseline.

5.3 Combination of Words' Starting or Ending Methods with Affix Morphemes

The experiments were redone combining words' starting or ending methods and Malay affix morphemes for tagging unknown words. Three factors influenced the results, i.e. the number of training iterations, the maximum predefined length of characters used in a words' starting or ending methods and the number of joint "word-tag" and word types used in the successive abstraction smoothing. Words that contained affixes were ignored when counting the total number of joint "word-tag" that shared the same sequence of words' starting or ending used in (12) and the total number of word types that shared the same sequence of words' starting or ending used in (13). The experiment was repeated for each maximum predefined length of characters, ranging from one (1) to twelve (12) characters, for a words' starting and ending methods with various numbers of iterations, ranging from one (1) to ten (10). For each iteration, four sets of probabilities were prepared:

(a) The probability of a tag t_i given the first m letters, as formulated in (8).
(b) The probability of a tag t_i given the last m letters (similar to (a) but using the reverse order of characters).
(c) The probability of a tag t_i given the first m letters, ignoring the affixed words.
(d) The probability of a tag t_i given the last m letters, ignoring the affixed words.

In the tagging process, if the tagger finds an unknown word, the tagger checks whether the word contains an affix morpheme. If the unknown word contains an affix morpheme, the emission is replaced by a value proportionate to the marginal distribution of tags. If it does not, the emission is replaced by either the probability in (a) to (d) above. The results were shown in the last four rows in Table 4 (row 6-9). The best combination was found using affix morpheme POS guessing with emission replaced by a value proportionate to the marginal distribution of tags and the words' ending information, which was the

Table 4 Tagging Performance

	Methods	Maximum predefined length of characters	Training iterations	Overall	Seen words			Unseen words
					Exist in dictionary		Not exist in dictionary	
					Unique tag	Ambiguous tags		
1	Baseline - POS guesser algorithm using affix morphemes	-	-	38.50	42.30	7.08	40.31	30.10
2	HMM-Viterbi with words' starting	4	2	81.81	92.00	75.78	38.97	33.42
3	HMM-Viterbi with words' ending	8	3	81.71	92.00	75.83	38.33	32.22
4	HMM-Viterbi with morpheme - uniform distribution to all possible tags	-	2	82.25	92.00	75.52	42.90	31.22
5	HMM-Viterbi with morpheme - marginal proportionate distribution of tags	-	2	82.28	92.00	76.04	42.52	31.94
6	HMM-Viterbi with combination of morphemes and words' starting	8	2	82.53	92.00	76.19	43.93	33.72
7	HMM-Viterbi with combination of morphemes and words' ending	5	2	82.59	92.00	76.14	44.56	33.40
8	HMM-Viterbi with combination of morphemes and words' starting with successive abstraction smoothing ignores words have affixes	1	2	82.58	92.00	76.13	44.32	34.20
9	HMM-Viterbi with combination of morphemes and words' ending with successive abstraction smoothing ignores words have affixes	6	2	**82.72**	92.00	**76.36**	**45.13**	**34.24**

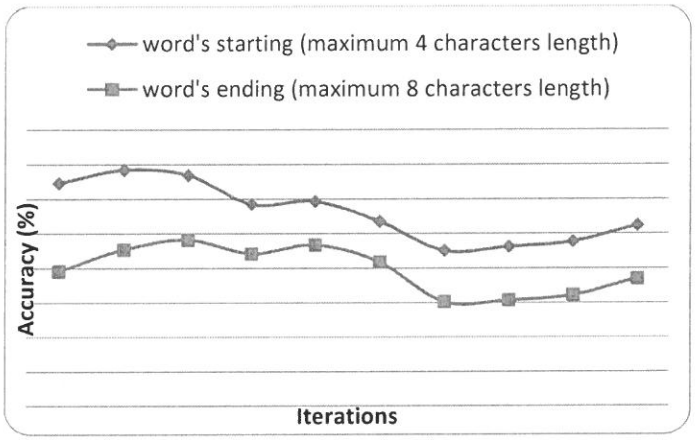

Fig. 5 The overall performance of HMM-Viterbi with character-based predictions over various training iterations

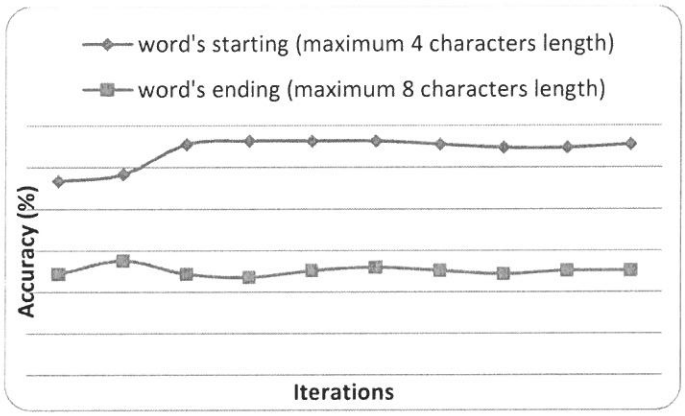

Fig. 6 The performance of tagging unseen words over various training iterations

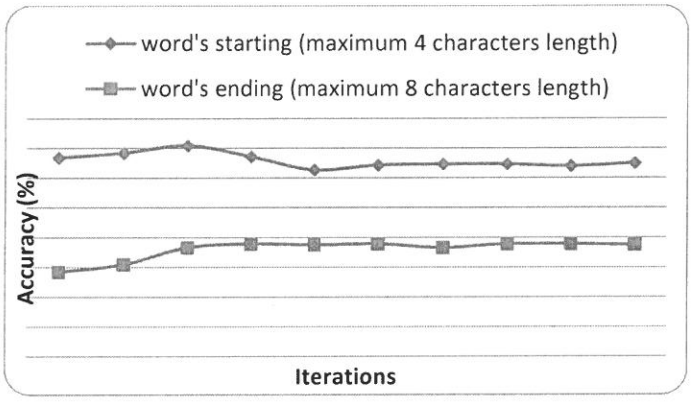

Fig. 7 The performance of tagging words not found in the dictionary over various training iterations

probability of a tag t_i given the six predefined length of letters with successive abstraction smoothing ignored affixed words (using the probability in (d)). The best combination was found using affix morpheme POS guessing with emission replaced by a value proportionate to the marginal distribution of tags and the

words' ending information, which was the probability of a tag t_i given the six predefined length of letters with successive abstraction smoothing ignored affixed words (using the probability in (d)). This combination performed overall tagging with 82.72% accuracy, the highest among all combinations, and was found to have high accuracy in guessing tags for words that do not exist in the dictionary; outperforming the baseline. Without a combination, the HMM-Viterbi using a words' starting information (maximum 4 characters of predefined length) was good for tagging unseen words, which outperformed the baseline.

6.0 DISCUSSION

A words' starting and ending predictions model implicitly includes Malay linguistics, which is affix information. It extends the paradigm of affixations in linguistic meaning. Malay affixes have some significant statistical distribution. The distribution of words containing circumfixes, prefixes or suffixes in the Malay language is almost consistent for any different corpus sizes. Fig. 8, Fig. 9 and Fig. 10 showed the distribution of affixes found in the training corpus (995,240 tokens) and test corpus (121,090 tokens). The test corpus had 17,818 tokens of unknown words identified as not listed in the dictionary, or 17.45% of the test corpus. From this number, 44.46% of words had affixes. Therefore, 45.13% of tagging accuracy for words not in the dictionary using the combination of a words' ending method with smoothing ignored affixed words and affix morpheme for POS guessing in Table 4 (row 9) was near to the percentage of words not listed in the dictionary with affixes (44.46%). Therefore, focusing on only derivative words with affixation, 97.13% were correctly tagged using a combination of words' ending with smoothing ignored affixed words and morpheme-based POS guessing. This percentage indicated that for affixed words, using morpheme-based POS guessing was quite effective.

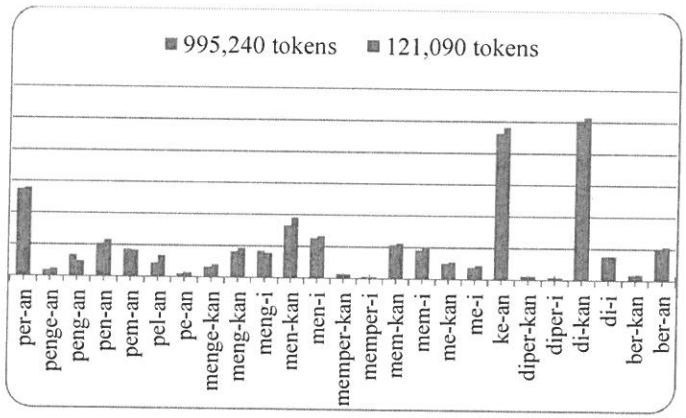

Fig. 8 The distribution of Malay circumfixes in two different corpuses

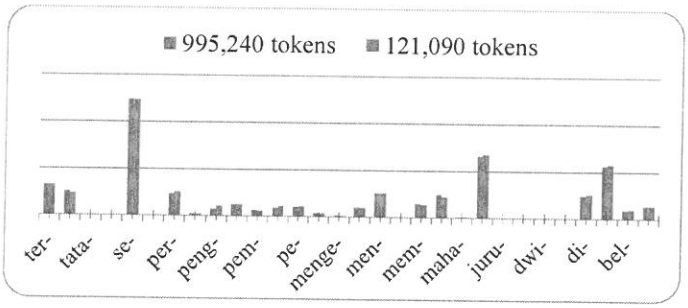

Fig. 9 The distribution of Malay prefixes in two different corpuses

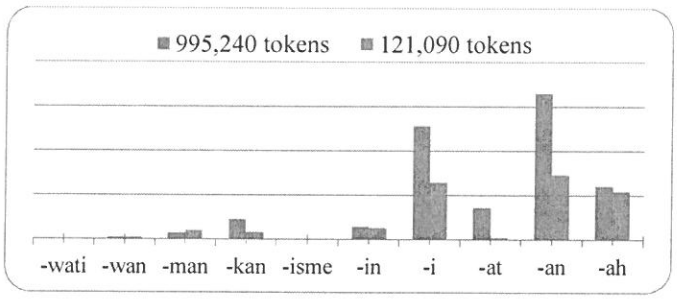

Fig. 10 The distribution of Malay suffixes in two different corpuses

The perfect guessing by affix morphemes was shown by circumfixes *meng-...-kan* and *memper-...-i* with error percentage of zero. However, if only reliable data is considered, which is data that occurred more than 100 times, the circumfix *di-...-kan* is the best affix morpheme for guessing because it was repeated 1,176 times (appeared in unknown words) in the tests corpus with tagging error as low as 5.36%. The bad guessing identified as unknown words containing the morpheme *ke-...-an* was with the error percentage of 28.08%, over 381 unknown words. The poor result was expected because the morpheme *ke-...-an* has ambiguous tags that are KN, KK and KA.

Table 5 displayed the significance of Malay circumfixes according to the lowest error rate when regarding morphemes used for guessing the unknown words' POS. This analysis was based on the result in Table 4 (row 5) in which unknown words were handled by Malay affix morphemes by replacing the unknown words' emissions by marginal proportionate distribution of tags. According to observations, Malay words that contain the circumfix *di-...-kan* were rarely listed as entries in the dictionary. The circumfix *di-...-kan* is used to derive passive verbs, which is quite similar to suffix *...-ed* in English, for indicating the past tense; also rarely listed in dictionaries. Therefore, our method to penalise the emission probabilities of unknown words using Malay affix morphemes was effective for certain morphemes.

Table 5 The best circumfixes for guessing unknown words' POS

Circumfixes	Percentage of errors
di-...-kan	5.36%
per-...-an	5.56%
di-...-i	8.13%
mem-...-kan	11.27%
mem-...-i	14.83%
pen-...-an	17.30%
ber-...-an	21.09%
pem-...-an	21.95%

Table 6 The best prefixes for guessing unknown words' POS

Prefixes	Percentage of errors
men-	4.63%
se-	7.45%
di-	9.29%
ber-	20.20%
peng-	21.15%
ter-	27.02%
ke-	28.27%

Table 6 showed the significance of Malay prefixes according to the lowest error rate when the regarding morphemes were used for guessing the unknown words' POS. This analysis was also based on the result in Table 4 (row 5). According to the consideration that only some amount of words are considered as reliable data (words that occur more than 100 times), the prefix *di-* showed the best guessing with an error rate of 9.29% from 990 words. The bad guessing was identified as unknown words that contained prefix *be-...* with an error rate of 66.67% from 105 words. The reason for this was that the prefix *be-...* clashed with Malay words that originally began with *be-...* (such as *begitulah*, *benar-benar*, *benihnya*, etc.) in a way that the words became unknown because their word form orthographically changed after adding particle *lah*, clitic *nya* or hyphen.

According to the consideration that only some amount of words are reliable data (words that occur more than 100 times), the suffix *...-i* showed the best guessing with an error rate of 15.76% from 590 words. The reason for this high error rate was because the morpheme *...-i* clashed with Malay words that originally end up with letter *i*, such as *ahli-ahli*, *saksi-saksi*, *Hilmi*, *Fahmi*, *koboi, etc.* In general, guessing the POS using Malay suffixes gave inaccurate results; for example, the suffix *...-an* had successfully guessed only half of the 644 words (error rate 49.69%). The other prefixes showed an error rate higher than 50%. The original Malay prefixes in the rules were only *-an*, *-kan* and *-i*; the others were from use in foreign words, such as *...-in*, *...-ah*, *...-at* from Arabic and *...-isme* from English.

7.0 CONCLUSIONS

A Malay POS tagger was developed based on the unsupervised Hidden Markov Model (HMM). The challenge in unsupervised HMM training was that training is dependent on an untagged corpus and the only supervised portion to limit possible word tags was the dictionary. In this study, a published dictionary was used. The dictionary did not include all words found in the corpus, especially derivative words such as passive verbs and derivative nouns. Therefore, the training outcome had a problem with unknown words, not just words not found in the corpus, but also with words that appeared but were not listed in the dictionary. Hence, properly mapping possible tags was challenging. Effort was put into looking at the exact morphemes of prefixes, suffixes and circumfixes in the agglutinative Malay language. When tagging a new sentence, words in the sentence identified as not listed in the dictionary were assigned with probable tags based on linguistically meaningful affixes, as defined in morphological rules through the morpheme-based POS guessing algorithm.

HMM-Viterbi tagging with a words' starting information was better than using a words' ending information for guessing unknown words' POS. A good overall accuracy was achieved using a words' starting information, with the need to check a maximum of four characters in predefined length (81.81%) compared to using a words' ending information, where a maximum of six characters predefined length (81.71%) was necessary.

The overall performance of HMM-Viterbi tagging with morpheme-based POS guessing showed that the unknown word emissions replaced by the value proportionate to marginal distribution of possible tags of unknown words (82.28%) was better than the words' emission replaced by the same distribution of all possible tags of the unknown word (82.25%). However, emissions replaced by the same distribution of all possible tags was good for tagging words not listed in the dictionary (42.90%). On the other hand, emissions replaced by a value proportionate to the marginal distribution were good for tagging unseen words (31.94%). HMM-Viterbi tagging with morpheme-based POS guessing (good overall accuracy was 82.28%) was better than HMM-Viterbi tagging with a words' starting information prediction (good overall accuracy was 81.81%). Therefore, tagging unknown words identified as not existing in the dictionary was better with the assistance of morpheme-based POS guessing (42.52%).

The best combination was found to be using morpheme-based POS guessing with emissions replaced by a value proportionate to the marginal distribution of tags and a words' ending information, given the six predefined length of characters, ignoring affixed words in successive smoothing. This combination showed an overall tagging accuracy of 82.72%, the highest among all, also proved good at guessing tags for words

not in the dictionary; which outperformed the baseline. Without using the combination, HMM-Viterbi using a words' starting information (maximum of 4 characters predefined length) was good for tagging unseen words, and outperformed the baseline and others.

REFERENCES

Abdullah, H. (2006). *Morfologi siri pengajaran dan pembelajaran bahasa Melayu.* PTS Professional, Kuala Lumpur.

Abdullah, H., Seri Lanang, J. R., Razali, A. & Zulkifli, O. (2006). *Sintaksis siri pengajaran dan pembelajaran bahasa Melayu.* PTS Professional, Kuala Lumpur.

Arbak, O. (2005). *Kamus komprehensif bahasa Melayu.* Oxford Fajar, Shah Alam.

Asmah, O. (2009). *Nahu Melayu mutakhir.* Dewan Bahasa dan Pustaka, Kuala Lumpur.

Azimizadeh, A., Mehdi, M. A. & Quchani, S. R. (2008). Persian Part of Speech Tagger Based on Hidden Markov Model. *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, pp. 121-128.

Banko, M. & Moore,R. C. (2004). Part of Speech Tagging in Context. *Computational Linguistics Association for Computational Linguistics*, pp.556-561.

Brants, T. (1995). Tagset reduction without information loss. *ACL*, pp. 287-289.

Brants, T. (2000). TnT: a statistical part-of-speech tagger. *Applied Natural Language Processing*, pp. 224-231.

Cucerzan, S. & Yarowsky, D. (2000). Language independent, minimally supervised induction of lexical probabilities. *Association for Computational Linguistics*, pp. 270–277.

Dandapat, S. (2009). *Part-of-Speech Tagging for Bengali.* Master Thesis, Department of Computer Science and Engineering, Indian Institute of Technology.

Das, D. & Petrov, S. (2011). Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections. *Association for Computational Linguistics: Human Language Technologies*, 1:600-609.

Dienes, P. & Oravecz, C. (2000). Bottom-up tagset design from maximally reduced tagset. *COLING Workshop on Linguistically Interpreted Corpora*, pp. 42-47.

Dominguez, M. A. & Infante-Lopez, G. (2008). Searching for part of speech tags that improve parsing models. *Advances in Natural Language Processing (ANLP)*, pp 126-137.

Garrette, D. & Baldridge, J. (2012). Type-Supervised Hidden Markov Models for Part-of-Speech Tagging with Incomplete Tag Dictionaries. *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 821–831.

Garrette, D. & Baldridge, J. (2013). Learning a Part-of-Speech Tagger from Two Hours of Annotation. *NAACL-HLT*, pp. 138–147.

Garrette, D., Mielens, J. & Baldridge, J. (2013). Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages. *Association for Computational Linguistics*, pp. 583–592.

Giesbrecht, E. & Evert, S. (2009). Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. *Web as Corpus Workshop (WAC5)*, pp. 27-35.

Goldberg, Y., Adler, M. & Elhadad, M. (2008). EM can find pretty good HMM pos-taggers (when given a good start. *Association for Computational Linguistic*, pp. 746-754.

Güngör, T. (2010). Part-of-speech Tagging. In Indurkhya, N. & Damerau, F. J. (Eds.), *Handbook of Natural Language Processing*, Taylor and Francis, pp. 205-235.

Haghighi, A. & Klein, D. (2006). Prototype-Driven Learning for Sequence Models, *HLT/NAACL*, pp. 320-327.

Hall, J. (2003). *A Probabilistic Part-of-Speech Tagger with Suffix Probabilities.* Master Thesis, School of Mathematics and Systems Engineering, Växjö University.

Hasan, K. S. & Ng, V. (2009). Weakly Supervised Part-of-Speech Tagging for Morphologically-Rich, Resource-Scarce Languages. *European Chapter of the ACL*, pp. 363–371.

Hawkins, M. J. (2008). *Kamus dwibahasa Bahasa Inggeris – Bahasa Malaysia.* Oxford Fajar, Shah Alam.

Hock, O. Y. (2009). *Kamus Dwibahasa.* Paerson Longman, Petaling Jaya.

Kupiec,J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6(6):225-242.

Marcus, P. M., Santorini, B. & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 2(19):313-330.

Merialdo, B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, 2(20):155-171.

Miller, J. E., Torii, M. & Vijay-Shanker, K. (2007). Adaptation of POS tagging for multiple BioMedical domains. *Biological, Translational, and Clinical Language Processing (BioNLP)*, pp. 179-180.

Mohseni, M., Motalebi, Minaei-bidgoli, H., B. & Shokrollahi-far, M. (2008). A Farsi Part of Speech Tagger Based on Markov Model. *ACM Symposium on Applied Computing*, pp. 1588-1589.

Nik Safiah, K., Farid, O.M., Hashim, M. & Abdul Hamid, M. (2010). *Tatabahasa dewan edisi ketiga*. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Petrov, S., Das, D. & McDonald, R. (2012). A Universal Part-of-Speech Tagset. *Language Resources and Evaluation (LREC'12)*, pp. 2089-2096.

Ravi, S. & Knight, K. (2009). Minimized Models for Unsupervised Part-of-Speech Tagging. *Natural Language Processing of the AFNLP*, pp. 504-512.

Samuelsson, C. (1996). Handling sparse data by successive abstraction. *Computational linguistics*, 2:895-900.

Schröder, I. (2002). *A Case Study in Part-of-Speech Tagging Using the ICOPOST Toolkit*. Technical report FBI-HH-M-314/02, Department of Computer Science, University of Hamburg.

Subramanya, A., Petrov, S., & Pereira, F. (2010). Efficient Graph-Based Semi-Supervised Learning of Structured Tagging Models. *Empirical Methods in Natural Language Processing*, pp. 167–176.

Täckström, O., Das, D., Petrov, S., McDonald, R. & Nivre, J. (2013). Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *Transactions of the Association for Computational Linguistics*, 1:1-12.

Toutanova, K. & Johnson, M. (2007). A Bayesian LDA-based Model for Semisupervised Part-of-Speech Tagging. *Neural Information Processing Systems (NIPS)*, pp. 1521–1528.

Vadas, D. & Curran, J. (2005). Tagging unknown words with raw text features. *Australasian Language Technology Workshop (ALTW)*, pp. 32–39.