http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 4 No. 1, June 2015: 11 - 23
e-ISSN: 2289-2192

MALAY PART OF SPEECH TAGGER: A COMPARATIVE STUDY ON TAGGING TOOLS

HASSAN MOHAMED
NAZLIA OMAR
MOHD. JUZAIDDIN AB. AZIZ

ABSTRACT

Malay language is an agglutinative language which rich morphology. Affixation to a root word *is the most common morphological processes used to derive a new word for other meaning that would affect the change in their part of speech (POS). Malay annotated corpus is not freely available, so there is no publication report on the comparison of the performance of POS tagging using Hidden Markov Model (HMM), Maximum Entropy (ME) and Support Vector Machine (SVM), especially to look into the effect of Malay morphology for tagging unknown words. This paper aims to present the evaluation of TnT using HMM, MaxEnt using ME and SVMTool using SVM. In order to train and test such methods in tagging Malay language, efforts has been taken to annotate the Malay corpus in health domain. Modifications has been done to TnT to fit in prefix and circumfix features. The results of the experiments shows that SVMTool outperforms TnT and MaxEnt for overall accuracy (99.23% for SVMTool, 94% for TnT and 96% for Maxent) and tagging unknown words accuracy (96.78% for SVMTool, 67% for TnT and 86.23% for MaxEnt). MaxEnt outperforms TnT for the overall accuracy and tagging unknown words. As the tagging accuracy of SVMTool to unknown word succeeds 96.78%, it would be the best tool for tagging Malay language for a specific domain.*

Keywords: Malay POS tagger, Malay morphemes, unknown words

ABSTRAK

Bahasa Melayu merupakan bahasa aglutinatif yang kaya dengan morfologi bagi menerbit perkataan dengan makna selain daripada kata akar yang memberi kesan kepada perubahan golongan katanya. Korpus beranotasi Bahasa Melayu sukar didapati lantas belum ada penerbitan tentang perbandingan prestasi penandaan golongan kata (GK) mengguna kaedah Model Markov Tersembunyi (MMT), Entropi Maksimum (EM) dan Mesin Vektor Sokongan (MVS), terutamanya bagi melihat kesan morfologi Bahasa Melayu ke atas penandaan GK bagi perkataan anu. Kertas ini bertujuan membentang penilaian ketiga-tiga kaedah tersebut ke atas Bahasa Melayu. Tiga alatan penanda GK diguna yakni *TnT* mewakili MMT, *MaxEnt* mewakili EM dan *SVMTool* mewakili MVS. Bagi melengkapi latihan dan ujian bagi ketiga-tiga alatan tersebut, usaha menganotasi korpus Bahasa Melayu bagi domain kesihatan dilakukan. Alatan *TnT* diubah suai untuk memasukkan fitur imbuhan awalan serta apitan. Keputusan bagi seluruh eksperimen menunjukkan prestasi *SVMTool* mengatasi *TnT* dan *MaxEnt* bagi kejituan keseluruhan (99.23% untuk *SVMTool*, 94% untuk *TnT* dan 96% untuk *MaxEnt*) serta kejituan penandaan perkataan anu (96.78% untuk *SVMTool*, 67% untuk *TnT* dan 86.23% untuk *MaxEnt*). Keupayaan *MaxEnt* pula mengatasi *TnT* bagi kejituan keseluruhan serta kejituan penandaan perkataan anu. Ketepatan penandaan perkataan anu sebanyak 96.78% oleh *SVMTool*, menjadikan alatan tersebut sebagai yang tebaik pada ketika ini dalam penandaan GK Bahasa Melayu bagi domain spesifik.

Kata kunci: Penanda Golongan Kata, Imbuhan, Perkataan Anu

PENGENALAN

Kajian susastera mendapati beberapa pendekatan penandaan golongan kata (GK) diguna seperti pendekatan berasaskan statistik, teknik pembelajaran mesin dan juga pendekatan petua. Contoh penanda GK mengguna pendekatan statistik seperti Markov Tersembunyi (MMT)

(Weischedel et al., 1993; Brants, 2000), Entropi Maksimum (EM) (Ratnaparkhi, 1996) serta pembelajaran berasaskan transformasi (Brill, 1995). Manakala contoh yang mengguna teknik pembelajaran mesin seperti pembelajaran berasaskan ingatan (Daelemans et al., 1996), pepohon keputusan (Màrquez & Rodríguez, 1997), AdaBoost (Abney et al., 1999), dan Mesin Vektor Sokongan (MVS) (Nakagawa et al., 2001). Kebanyakan penanda tersebut dinilai ke atas korpus Bahasa Inggeris *Wall Street Journal* (WSJ) yang dianotasi dengan set tanda GK *Penn Treebank* dan leksikon dibina secara langsung daripada korpus tersebut. Pada penghujung tahun 90-an, terdapat konsensus bagi menyatakan tahap pencapaian penandaan GK Bahasa Inggeris adalah di antara 96.4% hingga 96.7% (Giménez & Màrquez, 2003; Jurafsky & Martin, 2009; Manning & Schutze, 1999), walaupun terdapat sedikit perbezaan keadaan dalam penilaian yang dilakukan (seperti set eksperimen yang mengguna korpus yang berbeza atau set tanda yang berbeza).

Penanda yang paling berjaya dan popular dalam komuniti Pemprosesan Bahasa Tabii (PBT) ialah penanda *TnT* berasaskan MMT (Brants, 2000), pembelajaran berasaskan transformasi (*TBL tagger*) (Brill, 1995), dan beberapa varian pendekatan Entropi Maksimum (EM) (Ratnaparkhi, 1996) serta *SVMTool* yang berasaskan mesin vektor sokongan (Giménez & Màrquez, 2004). *TnT* salah satu contoh penanda yang praktikal untuk aplikasi PBT kerana tersedia secara terbuka (*open source*), ringkas dan mudah diguna serta tidak bersandar bahasa. *TnT* juga cekap kerana latihan yang dijalankan mengguna korpus satu juta perkataan hanya mengambil masa beberapa saat serta boleh menanda ribuan perkataan daripada teks baharu dalam sesaat (Schröder, 2002).

Kewujudan korpus beranotasi WSJ memboleh penyelidik memperbaiki keputusan penandaan GK secara berterusan. Antaranya ialah Thede dan Haper (1999) serta Lee et al. (2000) yang memboleh model MMT lebih berkeupayaan dan kompleks. Toutanova dan Manning (2000) pula memperkaya set fitur bagi penanda EM. Nakagawa et al. (2001) mengguna teknik pembelajaran SVM dengan latihan berasaskan Perseptron Undian bagi model EM (Collins, 2002). Penanda yang kompleks ini meningkat ketepatannya sehingga 97.1% yang diuji mengguna korpus WSJ. Brill dan Wu (1998), Halteren et al. (1998) serta Màrquez et al. (1999) mengemuka gabungan beberapa penanda yang sedia ada ke dalam skim undian alternatif dan mencapai ketepatan yang baik (kira-kira 97.2%).

Bagi kedudukan Bahasa Melayu, korpus beranotasi sukar didapati. Justeru, sepanjang pengetahuan penulis, belum ada penerbitan tentang perbandingan prestasi penandaan GK mengguna kaedah MMT, EM dan MVS ke atas Bahasa Melayu. Kertas ini bertujuan membentang penilaian alat penanda GK berpenyelia mengguna data Bahasa Melayu. Tiga alatan penanda GK yang diguna yakni *TnT* (Brants, 2000) yang mengguna MMT, *MaxEnt* (Ratnaparkhi, 1996) yang mengguna Entropi Maksimum (EM) dan *SVMTool* (Giménez & Màrquez 2004) yang mengguna mesin vektor sokongan. Memandangkan ketiga-tiga alatan tersebut memerlu korpus bertanda, usaha menganotasi korpus dilakukan. Latihan dan ujian bagi ketiga-tiga alatan tersebut dilakukan ke atas korpus tersebut bagi mendapatkan perbandingan prestasi.

PENYEDIAAN KORPUS BERTANDA BAHASA MELAYU

Sebelum melakukan anotasi ke atas korpus, set tanda GK dikenal pasti terlebih dahulu. Set golongan kata (GK) Bahasa Melayu dilihat pelbagai yang bergantung kepada penggunaan atau kecenderungan para sarjana seperti yang terdapat dalam pelbagai kamus (Arbak, 2005; Hock, 2009; Hawkins, 2008) buku teks (Abdullah et al., 2006; Nik Safiah et al., 2010) atau dicipta oleh penyelidik pengkomputeran linguistik (Ranaivo-Malancon, 2008). Sebagai kajian awal, analisis terhadap set GK Penn Treebank dilakukan. Analisis mendapati tidak semua tanda GK *Penn Treebank* memenuhi ciri-ciri Bahasa Melayu selain daripada terdapat kekurangan yang

lain. Sebagai contoh, dalam set tanda Penn Treebank, tanda kelas yang tidak relevan bagi Bahasa Melayu adalah VBZ (*verb, third person singular present*), VBP (*verb, non-third person singular present*), VBD (*verb, past tense*), VBN (*verb, past participle*) serta VBG (*verb, gerund or present participle*) kerana asas bagi kelas tersebut adalah kata kerja yang dikaitkan dengan pelaku serta masa perlakuan itu dibuat, sedangkan dalam Bahasa Melayu kata kerja tidak dikaitkan secara jelas dengan pelakunya sama ada jamak atau tunggal serta masa perbuatan itu dilakukan. Kekurangan yang jelas dalam set tanda *Penn Treebank* adalah ketiadaan tanda kelas penjodoh bilangan yang wujud dalam Bahasa Melayu.

Terdapat set tanda GK Dewan Bahasa dan Pustaka (DBP) yang ditakrif oleh Knowles dan Zuraida (2006). Set ini dianggap boleh diterima kerana DBP adalah badan yang berautoriti berkaitan Bahasa Melayu di Malaysia. DBP mempunyai 18 kelas kata yang utama. Set GK yang diguna dalam kamus Hock (2009) dan Arbak (2005) adalah serupa dengan kelas utama DBP. Jadual 1 menunjukkan perbandingan kelas utama DBP dengan set GK Bahasa Melayu yang diguna dalam kamus dwibahasa Melayu - Bahasa Inggeris Hock (2009). Oleh itu, set GK dalam kamus dwibahasa Hock (2009) diguna dalam kajian ini dengan adaptasi daripada set tanda DBP. Kamus tersebut mengandungi 576 perkataan yang mempunyai tanda GK kabur atau 96.41% perkataan mempunyai tanda GK yang unik.

JADUAL 1. Perbandingan set tanda GK DBP dan set tanda GK dalam kamus Hock (2009)

Set tanda GK DBP	Set tanda GK dalam kamus Hock (2009)	Penerangan
N	KN	Kata Nama
K	KK	Kata Kerja
S	ADJ	Kata Adjektif
I	KSN	Kata Sendi
-	KB	Kata Bantu
	KG	Kata Ganti Nama
H	KH	Kata Hubung
A	ADV	Adverba
-	SR	Kata Seru
T	KT	Kata Tanya
B	KBIL	Kata Bilangan
-	KPM	Kata Pemeri
-	KP	Kata Perintah
-	KAR	Kata Arah
W	PW	Penanda Wacana
-	AWL	Awal
-	KEP	Kependekan
#	-	Nombor
$	-	Simbol Mata wang
%	-	Simbol Huruf
D	-	Kata Deiksis
G	-	Kata Nama Khas
L	-	Senarai
P	-	Kata Penanda
X	-	Kata Nafi
Z	-	Kata Pinjam

Terdapat beberapa tanda dalam kamus Hock (2009) digugur dan ditambah dengan tanda lain. Misalnya, tanda AWL dan KEP digugur kerana tanda tersebut adalah bukan kelas kata dari segi bahasa. Klitik adalah penting dalam Bahasa Melayu kerana memberi fungsi sebagai kata ganti nama seperti *nya*, *mu* dan *ku.* Selain daripada klitik terdapat juga partikel seperti *lah* dan *kah.* Klitik dan partikel ini mesti dikendali dengan baik. Bagaimanapun, pada peringkat awal dan bagi tujuan kemudahan, hanya klitik *nya* dan partikel *lah* sahaja yang dikendali. Ini

kerana korpus yang hendak dianotasi diambil dari artikel surat khabar yang mengandungi banyak *nya* dan *lah* dan tidak bertemu klitik yang lain seperti *mu* dan *ku*.

Perkataan yang dijerait dengan klitik atau partikel dipisah kepada dua token. Contohnya *terjejasnya* menjadi *terjejas* dan *nya*. Tanda @KG ditambah untuk menanda klitik, manakala tanda #E untuk menandai partikel. Selain isu klitik dan partikel, tanda KNF diguna bagi menandai perkataan nafi seperti *tidak* dan *bukan*, tanda KNK untuk menandai kata nama khas, SEN untuk menandai mana-mana nombor senarai, dan seterusnya SYM untuk menandai apa-apa simbol termasuk tanda bacaan. Jadual 2 menyenarai keseluruhan tanda GK yang diguna untuk anotasi yang berjumlah sebanyak 21 tanda.

JADUAL 2. Senarai Set Tanda yang digunakan dalam korpus

Set Tanda	Penerangan
KN	Kata Nama
KK	Kata Kerja
ADJ	Kata Adjektif
KSN	Kata Sendi
KB	Kata Bantu
KG	Kata Ganti Nama
KH	Kata Hubung
ADV	Adverba
KT	Kata Tanya
KBIL	Kata Bilangan
KPM	Kata Pemeri
KP	Kata Perintah
KAR	Kata Arah
PW	Penanda Wacana
KEP	Kependekan
#E	Partikel
@KG	Klitik
KNF	Kata Nafi
KNK	Kata Nama Khas
SEN	Senarai Nombor
SYM	Sebarang simbol atau tanda bacaan

Penyediaan korpus dibuat secara semi-automatik. Perkataan dalam korpus yang ada dalam entri kamus ditandai secara automatik melalui jadual carian. Perkataan yang gagal ditanda dengan cara ini kerana ketiadaan entri atau perkataan yang mempunyai tanda lebih daripada satu tanda (tidak unik) akan diedit secara manual. Proses mengedit korpus Bahasa Melayu adalah berdasarkan rumus morfologi Nik Safiah et al. (2010). Bahasa Melayu merupakan bahasa aglutinatif yang kaya dengan morfologi. Pengimbuhan adalah proses morfologi yang paling biasa diguna yang terdiri daripada tiga jenis imbuhan iaitu awalan, akhiran dan juga apitan. Menambah imbuhan kepada kata akar dapat mengubah makna asalnya yang juga memberi kesan kepada perubahan golongan katanya (Abdullah et al., 2006; Nik Safiah et al. 2010).

Jadual 3 menunjukkan golongan kata yang berpotensi selepas pengimbuhan. Pengimbuhan awalan, akhiran atau apitan yang mengubah kategori kepada kata nama, kata akarnya adalah kata nama atau kata kerja. Pengimbuhan awalan, akhiran atau apitan yang mengubah kategori kepada kata kerja, kata akarnya adalah daripada kata nama atau kata kerja. Pengimbuhan awalan atau apitan yang mengubah kategori kepada kata adjektif, kata akarnya adalah kata adjektif. Korpus yang dianotasi mengandungi kira-kira 18,135 token dengan 1,381 jenis perkataan yang mempunyai tanda kabur. Setelah korpus siap, semua fitur yang dikehendaki oleh alatan *TnT*, *MaxEnt* dan *SVMTool* diekstrak daripada korpus tersebut.

JADUAL 3. Imbuhan awalan, akhiran dan apitan Bahasa Melayu dengan kemungkinan kategori perkataan

Kata Nama (KN)	Kata Kerja (KK)	Kata Adjektif (KA)
ke-…-an, pe-…-an, peN-…-an, peR-…-an	beR-…-an, beR-…-kan, di-…-i, di-…-kan, dipeR-…-i, dipeR-…-kan, ke-…-an, meN-…-i, meN-…-kan, mempeR-…-i, mempeR-…-kan	ke-…-an
juru-, ke-, peN-, pe-, peR- -an	beR-, di-, dipeR-, meN-, mempeR-, teR- -i, -kan	ter-, se-

MODEL MARKOV TERSEMBUNYI (MMT)

Terdapat tiga kaedah utama dalam penandaan yang dibuat ke atas Bahasa Melayu iaitu mengguna Model Markov Tersembunyi (MMT), Entropi Maksimum (EM) dan Mesin Vektor Sokongan (MVS). Penandaan MMT dibuat melalui alatan *TnT*, penandaan EM dibuat melalui *MaxEnt* dan penandaan MVS dibuat melalui *SVMTool*. *TnT* dan *MaxEnt* boleh dimuat turun melalui pakej ACOPOST http://acopost.sourceforge.net/, manakala *SVMTool* boleh dimuat turun melalui http://www.lsi.upc.es/~nlp/SVMTool/. Terdapat pengubahsuaian dalam alatan *TnT* yang dibuat bertujuan bagi memasukkan fitur imbuhan awalan serta apitan.

TnT mengguna model Markov tertib kedua untuk penandaan golongan kata. Keadaan dalam model mewakili tanda manakala lepasan mewakili perkataan. Kebarangkalian peralihan bergantung kepada perpindahan keadaan yakni pasangan tanda GK. Kebarangkalian pelepasan hanya bergantung kepada kategori yang paling terkini (tanda GK semasa).

Kebarangkalian peralihan dan kebarangkalian pelepasan dianggar daripada korpus beranotasi. Mengguna kebarangkalian kebolehjadian maksimum $\hat{P}$ yang diperoleh daripada frekuensi relatif, maka nilai monogram $\hat{P}(t_3)$, dwigram $\hat{P}(t_3|t_2)$, trigram $\hat{P}(t_3|t_1,t_2)$ dan leksikon $\hat{P}(w_i|t_i)$ ditentukan. Pelicinan kebarangkalian trigram $P(t_3|t_1,t_2)$ dilakukan dengan mengguna interpolasi linear hapus (*deleted linear interpolation*).

Penanda GK *TnT* meramal perkataan anu bagi bahasa fleksi melalui analisis akhiran perkataan dengan kebarangkalian tanda ditentu mengikut urutan huruf yang membentuk akhiran perkataan itu. Istilah akhiran perkataan yang diguna dalam *TnT* bermaksud "urutan terakhir aksara-perkataan" yang tidak semestinya imbuhan sebenar dari segi bahasa.

Taburan kebarangkalian untuk akhiran tertentu dihasil daripada semua perkataan dalam set latihan yang berkongsi akhiran yang sama dengan panjang maksimum yang ditetapkan. Kebarangkalian ini dilicin menurut kaedah abstraksi berturut-turut (Samuelsson, 1996) yang mengira kebarangkalian bersyarat tanda t diberi m huruf daripada n huruf akhiran perkataan dengan $m \leq n$.

Penanda GK *TnT* diubah suai untuk mengendali perkataan anu Bahasa Melayu dengan awalan perkataan. Kebarangkalian dilicinkan oleh abstraksi berturut-turut yang mengira kebarangkalian bersyarat tanda t diberi m huruf daripada n huruf awalan perkataan yakni $P(t|l_1,\ldots,l_m)$ dengan $m \leq n$. Pelicinan mengguna nilai $P(t|l_1,\ldots l_m)$, $P(t|l_1,\ldots l_{m-1}),\ldots,$ $P(t)$ yang didapati secara rekursif melalui pengiraan berikut:

$$P(t|l_1,\ldots l_{n+i}) = \frac{\hat{P}(t|l_1,\ldots l_{n+i}) + \Theta_i P(t|l_1,\ldots l_{n+i-1})}{1 + \Theta_i}$$

bagi setiap $i = 1 \ldots m$, dengan mengguna anggaran kebolehjadian maksimum $\hat{P}$ daripada frekuensi leksikon dalam korpus, pemberat Θ_i dan awalan nilai $P(t) = \hat{P}(t)$. Anggaran kebolehjadian maksimum bagi awalan perkataan dengan panjang i yang memberi tanda t dikira seperti berikut:

$$\hat{P}(t|l_1, \ldots l_n) = \frac{f(t, l_1, \ldots l_n)}{f(\, l_1, \ldots l_n)}$$

Idea untuk meramal tanda GK bagi perkataan anu dikembang melalui gabungan awalan dan akhiran perkataan. Maka, taburan kebarangkalian tercantum bagi awalan dan akhiran perlu dicari. Taburan kebarangkalian tercantum ini dijana daripada semua perkataan dalam set latihan yang berkongsi awalan dan akhiran yang sama daripada panjang maksimum yang ditetapkan.

Andaikan p dan s mewakili jujukan aksara yang masing-masing membentuk awalan dan akhiran. Mengguna petua rantai (*chain rule*), maka taburan kebarangkalian tercantum awalan dan akhiran diberi seperti berikut:

$$P(p, s) = P(p)P(s|p) \tag{1}$$

Bagaimanapun, penandaan mengguna model Markov memerlu kebarangkalian $P(p, s\,|t)$. Nilai ini boleh didapati dengan mengganti $P(p, s)$ dari persamaan (1) ke dalam penyongsangan Bayes seperti berikut:

$$\begin{aligned} P(p, s|t) &= \frac{P(t|p, s)\, P(p, s)}{P(t)} \\ &= \frac{P(t|p, s)\, P(p)P(s|p)}{P(t)} \end{aligned} \tag{2}$$

$P(t|p, s)$ dianggar dari korpus bertanda sebagai nisbah kekerapan tanda t bagi perkataan yang mengandungi gabungan awalan p dan akhiran s terhadap frekuensi perkataan yang mengandungi gabungan awalan p dan akhiran s seperti berikut:

$$P(t|p, s) = \frac{f(t, p, s)}{f(p, s)}$$

$P(s|p)$ dianggar dari korpus sebagai nisbah kekerapan perkataan dengan awalan p dan akhiran s terhadap kekerapan perkataan dengan awalan p seperti berikut:

$$P(s|p) = \frac{f(p, s)}{f(p)}$$

$P(p)$ dianggar sebagai nisbah kekerapan perkataan dengan awalan p terhadap frekuensi semua awalan N_{awalan}. Begitu juga dengan $P(t)$ yang boleh dianggar sebagai frekuensi tanda t terhadap jumlah frekuensi tanda yang diguna dalam korpus N_{tanda} (menyamai jumlah perkataan dalam korpus).

$$P(p) = \frac{f(p)}{N_{awalan}}$$

$$P(t) = \frac{f(t)}{N_{tanda}}$$

Dengan itu persamaan (2) dapat dikira dengan mudah seperti berikut:

$$P(p,s|t) = \frac{f(t,p,s)N_{tanda}}{f(t)N_{awalan}}$$

ENTROPI MAKSIMUM (EM)

Ratnaparkhi (1996) mengguna model entropi maksimum untuk membangun penanda GK *MaxEnt*. Model kebarangkalian ditakrif berdasarkan ruang $\mathcal{H} \times \mathcal{T}$, di mana $\mathcal{H}$ ialah 'sejarah' yang mengandungi set perkataan dan konteks tanda yang mungkin, dan $\mathcal{T}$ ialah set tanda yang dibenar berdasarkan sejarah. Dengan $\{w_1, ..., w_n\}$ mewakili urutan perkataan yang terdapat dalam data latihan dan $\{t_1, ..., t_n\}$ mewakili urutan tanda GK bagi perkataan tersebut, sejarah h_i dapat didefinisi. Latihan kepada parameter $\{\mu, \alpha_1, ..., \alpha_k\}$ berakhir apabila memaksimum kebolehjadian data latihan (*observations*).

Kebarangkalian tercantum sejarah h dan tanda t ditentu oleh parameter apabila fitur berkaitannya aktif, iaitu, bagi setiap α_j, $f_j(h,t) = 1$. Justeru, bagi setiap (h,t), sekurang-kurangnya satu fitur mengaktif perkataan atau tanda GK bagi sejarah h. Perkataan dan konteks tanda GK yang boleh menjadi fitur dimasukkan ke dalam takrif sejarah h_i. Sebagi contoh,

$$f_j(h_i,t_i) = \begin{cases} 1 & jika\ akhiran(w_i) = \text{"kan"}\ \&\ t_i = \text{KK} \\ 0 & selainnya \end{cases}$$

Fitur di atas menyumbang kepada kebarangkalian tercantum $p(h_i,t_i)$ apabila sesuatu perkataan w_i berakhir dengan "kan" dan apabila t_i = KK. Oleh itu parameter model α_j berfungsi sebagai "pemberat" untuk peramal konteks tertentu (dalam kes ini akhiran 'kan' diramal sebagai KK).

Model Entropi mengekstrak fitur dengan mengimbas setiap pasangan (h_i,t_i) dalam data latihan berpandukan 'templat' fitur yang diberi dalam Jadual 4. Pemberian nilai kepada pemboleh ubah X, Y, dan T dalam Jadual 4 ini diperoleh secara automatik dari data latihan.

JADUAL 4. Fitur sejarah semasa h_i

Syarat	Fitur	
w_i adalah perkataan kerap	$w_i = X$	$\& t_i = T$
w_i adalah perkataan jarang	X adalah awalan w_i, $\|X\| \leq 4$	$\& t_i = T$
	X adalah akhiran w_i, $\|X\| \leq 4$	$\& t_i = T$
	w_i mengandungi nombor	$\& t_i = T$
	Huruf awal w_i adalah huruf besar	$\& t_i = T$
$\forall w_i$	$t_{i-1} = X$	$\& t_i = T$
	$t_{i-2}t_{i-1} = XY$	$\& t_i = T$
	$w_{i-1} = X$	$\& t_i = T$
	$w_{i-2} = X$	$\& t_i = T$
	$w_{i+1} = X$	$\& t_i = T$
	$w_{i+2} = X$	$\& t_i = T$

Penjanaan fitur untuk penandaan perkataan anu adalah berdasarkan hipotesis bahawa perkataan 'jarang' dalam set latihan adalah serupa dengan perkataan anu dalam data ujian. Ramalan tanda GKnya bergantung kepada ortografi perkataan itu. Fitur perkataan jarang dalam

Jadual 4 akan melihat keserupaan ejaan perkataan bagi perkataan yang jarang dalam data latihan dan perkataan anu dalam data ujian. Sebagai contoh, Jadual 5 mengandungi petikan daripada data latihan "*berbual*/KK *kosong*/ADJ *sehingga*/KSN *berjam-jam*/KK *lama*/ADJ *nya*/@KG" dan fitur yang dijana semasa mengimbas (h_3, t_3) untuk perkataan semasa *sehingga* serta fitur yang dijana semasa mengimbas (h_4, t_4) untuk perkataan semasa *berjam-jam*. Perkataan *berjam-jam* berlaku hanya sekali dalam data latihan dan dengan itu dikelas sebagai 'jarang'.

JADUAL 5. Contoh ekstrak fitur bagi cebisan ayat

Sampel Data			Fitur (h_3, t_3) bagi perkataan semasa "*sehingga*"		Fitur (h_4, t_4) bagi perkataan semasa "*berjam-jam*"	
1	berbual	KK	w_i = sehingga	& t_i = KSN	w_{i-1} = *sehingga*	& t_i = KK
2	kosong	ADJ	w_{i-1} = kosong	& t_i = KSN	w_{i-2} = *kosong*	& t_i = KK
3	sehingga	KSN	w_{i-2} = berbual	& t_i = KSN	w_{i+1} = *lama*	& t_i = KK
	berjam-		w_{i+1} = berjam-jam	& t_i = KSN	w_{i+2} = *nya*	& t_i = KK
4	jam	KK	w_{i+2} = lama	& t_i = KSN	t_{i-1} = KSN	& t_i = KK
5	lama	ADJ	t_{i-1} = NNS	& t_i = KSN	$t_{i-2}t_{i-1}$ = ADJ KSN	& t_i = KK
6	nya	@KG	$t_{i-2}t_{i-1}$ = DT NNS	& t_i = KSN	awalan(w_i) = *b*	& t_i = KK
7	.	SYM			awalan(w_i) = *be*	& t_i = KK
					awalan(w_i) = *ber*	& t_i = KK
					awalan(w_i) = *berj*	& t_i = KK
					akhiran(w_i) = *m*	& t_i = KK
					akhiran(w_i) = *am*	& t_i = KK
					akhiran(w_i) = *jam*	& t_i = KK
					akhiran(w_i) = *-jam*	& t_i = KK

MESIN VEKTOR SOKONGAN (MVS)

Paradigma Mesin Vektor Sokongan (MVS) diguna untuk penandaan dalam Nakagawa et al. (2001) sebelum ini dengan tumpuan diberi kepada ramalan golongan kata bagi perkataan anu. Terdapat sedikit kelemahan kepada penanda ini iaitu kecekapannya yang rendah (yakni kelajuan penandaan hanya kira-kira 20 perkataan sesaat). Giménez dan Màrquez (2003) mengatasi keterbatasan ini dengan berurusan dengan kernel linear dalam pelarasan utama kerangka MVS yang mengambil kelebihan daripada contoh vektor yang sangat jarang-jarang. Penanda yang dihasilkan hampir setepat Nakagawa et al. (2001) tetapi 60 kali lebih cepat. Prototaip ini diatur cara dalam Perl yang dikenali sebagai *SVMTool* dan boleh dimuat turun dari http://www.lsi.upc.es/˜nlp/SVMTool/.

Penandaan perkataan menurut konteks adalah masalah klasifikasi multi-kelas. Memandangkan MVS adalah pengelas binari, membinarikan masalah harus dilakukan sebelum MVS boleh diguna. Giménez dan Màrquez (2003) mengguna pembinarian mudah yakni pembinarian *satu-bagi-setiap-kelas* (*one-per-class binarization*), iaitu SVM dilatih untuk setiap golongan kata bagi membeza antara contoh sesuatu kelas dengan yang lain. Apabila menandakan perkataan, tanda yang paling yakin menurut ramalan semua binari MVS akan dipilih.

Bagi pengekstrakan fitur untuk Bahasa Melayu, setiap contoh dikodifikasi berdasarkan konteks tempatan sesuatu perkataan. Tetingkap berpusat tujuh token dipertimbang untuk membentuk fitur binari. Oleh itu, pola asas dan n-gram dinilai misalnya, daripada contoh dalam Jadual 6, jika perkataan semasa adalah "*dengan*" maka "*sebelum_perkataan = masa*", "*dua_tanda_berturut-turut = KN_KSN*". Jadual 7 mengandungi senarai semua pola yang dipertimbang.

JADUAL 6. Sampel Ayat

Kedudukan	Perkataan	Tanda
-3	Mereka	KG
-2	membuang	KK
-1	masa	KN
0	dengan	KSN
+1	berbual	KK
+2	kosong	ADJ
+3	sehingga	KSN
+4	berjam-jam	KK
+5	lama	ADJ
+6	nya	@KG
+7	.	SYM

JADUAL 7. Pola fitur bagi perkataan semasa "dengan"

Fitur	Pola fitur	Contoh
Fitur Perkataan	w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}	w_{-3} = Mereka, w_{-2} = membuang, w_{-1} = masa, w_0 = dengan, w_{+1} = berbual, w_{+2} = kosong, w_{+3} = sehingga
Fitur GK	p_{-3}, p_{-2}, p_{-1}, p_0, p_{+1}, p_{+2}, p_{+3}	p_{-3} = KG, p_{-2} = KK, p_{-1} = KN, p_0 = KSN, p_{+1} = KK, p_{+2} = ADJ, p_{+3} = KSN
Kelas kabur	a_0, a_1, a_2, a_3	a_0 = KSN, a_1 = KK, a_2 = ADJ, a_3 = KSN
Dwigram perkataan	(w_{-2}, w_{-1}), (w_{-1}, w_{+1}), (w_{-1}, w_0), (w_0, w_{+1}), (w_{+1}, w_{+2})	(w_{-2}, w_{-1}) = (membuang, masa), (w_{-1}, w_{+1}) = (masa, berbual), (w_{-1}, w_0) = (masa, dengan), (w_0, w_{+1}) = (dengan, berbual), (w_{+1}, w_{+2}) = (berbual, kosong)
Dwigram GK	(p_{-2}, p_{-1}), (p_{-1}, a_{+1}), (a_{+1}, a_{+2})	(p_{-2}, p_{-1}) = (KK, KN), (p_{-1}, a_{+1}) = (KN, KK), (a_{+1}, a_{+2}) = (KK, ADJ)
Trigram perkataan	(w_{-2}, w_{-1}, w_0), (w_{-2}, w_{-1}, w_{+1}), (w_{-1}, w_0, w_{+1}), (w_{-1}, w_{+1}, w_{+2}), (w_0, w_{+1}, w_{+2})	(w_{-2}, w_{-1}, w_0) = (membuang, masa, dengan), (w_{-2}, w_{-1}, w_{+1}) = (membuang, masa, berbual), (w_{-1}, w_0, w_{+1}) = (masa, dengan, berbual), (w_{-1}, w_{+1}, w_{+2}) = (masa, berbual, kosong), (w_0, w_{+1}, w_{+2}) = (dengan, berbual, kosong)
Trigram GK	(p_{-2}, p_{-1}, a_0), (p_{-2}, p_{-1}, a_{+1}), (p_{-1}, a_0, a_{+1}), (p_{-1}, a_{+1}, a_{+2})	(p_{-2}, p_{-1}, a_0) = (KK, KN, KSN), (p_{-2}, p_{-1}, a_{+1}) = (KK, KN, KK), (p_{-1}, a_0, a_{+1}) = (KN, KSN, KK), (p_{-1}, a_{+1}, a_{+2}) = (KN, KK, ADJ)
Penutup ayat	'.', '?', '!'	
Awalan	s_1, s_1s_2, $s_1s_2s_3$, $s_1s_2s_3s_4$	s_1 = d, s_1s_2 = de, $s_1s_2s_3$ = den, $s_1s_2s_3s_4$= deng
Akhiran	s_n, $s_{n-1}s_n$, $s_{n-2}s_{n-1}s_n$, $s_{n-3}s_{n-2}s_{n-1}s_n$	s_n = n, $s_{n-1}s_n$ = an, $s_{n-2}s_{n-1}s_n$ = gan, $s_{n-3}s_{n-2}s_{n-1}s_n$ = ngan
Fitur binari	Pangkal perkataan huruf besar, Semua huruf besar, Semua huruf kecil, Ada tanda sempang, Token tidak bermula dengan huruf	Pangkal perkataan huruf besar = 0, Semua huruf besar = 0, Semua huruf kecil = 1, Ada tanda sempang = 0, Token tidak bermula dengan huruf = 0
Panjang perkataan	Integer	6

PENILAIAN

Kejituan penandaan diuji berdasarkan pengesahan silang. Korpus dibahagi kepada 10 petak. Satu petak dijadikan ujian manakala sembilan petak yang lain dicantum sebagai korpus latihan. Oleh itu korpus akan memperuntukkan 90% sebagai set latihan dan 10% ujian yang ditetapkan sedemikian hingga terdapat perkataan dalam data ujian yang tak kelihatan semasa latihan. Setiap keputusan diperoleh dengan mengulangi eksperimen 10 kali (Brants, 2000; Schröder, 2002) terhadap petak korpus yang berbeza dan mempuratanya untuk mendapat satu hasil. Peratusan purata perkataan anu adalah 15% daripada jumlah 1,840 purata token ujian. Saiz korpus termasuk latihan dan ujian adalah kira-kira 18,135 token. Korpus beranotasi yang digunakan dalam penilaian adalah daripada domain kesihatan yang didapati daripada artikel kesihatan Berita Harian *on-line*.

KEPUTUSAN

Keputusan bagi seluruh eksperimen ditunjukkan dalam Jadual 8. Keputusan menunjukkan prestasi *SVMTool* mengatasi *TnT* dan *MaxEnt* bagi kejituan keseluruhan (99.23% untuk

SVMTool, 94% untuk *TnT* dan 96% untuk *MaxEnt*) serta kejituan penandaan perkataan anu (96.78% untuk *SVMTool*, 67% untuk *TnT* dan 86.23% untuk *MaxEnt*). Keupayaan *MaxEnt* pula mengatasi *TnT* bagi kejituan keseluruhan serta kejituan penandaan perkataan anu. Ujikaji ini dijalankan ke atas korpus daripada domain kesihatan, yang menunjukkan, penandaan GK Bahasa Melayu bagi domain spesifik boleh mengguna alatan sedia ada sekiranya ada kesanggupan menyedia korpus bertanda dengan saiz yang dikira kecil (sekitar 15,000 hingga 18,000 token). Penandaan perkataan anu sebanyak 96.78% ketepatan oleh *SVMTool*, menandai bahawa isu penandaan GK Bahasa Melayu bagi domain spesifik buat ketika ini dianggap selesai.

JADUAL 8. Perbandingan kejituan penandaan GK bagi ketiga-tiga alatan

Korpus ujian	Kejituan keseluruhan			Kejituan GK untuk perkataan diketahui			Kejituan GK untuk perkataan anu		
	TnT	*MaxEnt*	*SVMTool*	*TnT*	*MaxEnt*	*SVMTool*	*TnT*	*MaxEnt*	*SVMTool*
Petak 1	91.81	94.82	98.60	98.39	98.63	99.71	66.82	85.16	95.13
Petak 2	93.72	95.82	99.14	98.51	98.80	99.82	67.95	86.07	96.43
Petak 3	95.33	96.76	99.57	98.94	99.21	99.95	68.71	87.00	98.09
Petak 4	95.73	96.90	99.65	99.00	99.26	99.98	69.83	87.06	98.38
Petak 5	94.80	96.53	99.48	98.72	99.13	99.92	68.46	86.69	97.72
Petak 6	93.07	95.44	99.11	98.38	98.64	99.91	66.88	85.51	96.08
Petak 7	94.49	96.24	99.29	98.56	98.91	99.88	67.92	86.26	96.56
Petak 8	93.16	95.57	99.05	98.40	98.68	99.80	67.03	85.98	96.23
Petak 9	93.32	95.66	99.10	98.44	98.74	99.83	67.37	86.09	96.36
Petak 10	94.57	96.32	99.33	98.62	98.96	99.89	68.06	86.51	96.79
Purata	94.00	96.00	99.23	98.60	98.90	99.87	67.90	86.23	96.78

PERBINCANGAN

Dalam penggunaan MMT trigram bagi penandaan Bahasa Melayu, prestasi terbaik untuk meramal GK perkataan anu melalui awalan perkataan adalah dengan menetapkan semaksimum tiga aksara panjang awalan tersebut. Jadual 9 menunjukkan prestasi penandaan mengguna MMT trigram dengan peramal GK perkataan anu mengguna huruf awalan perkataan dengan maksimum panjang awalan yang pelbagai.

Bagaimanapun, prestasi terbaik untuk meramal GK perkataan anu melalui akhiran perkataan adalah dengan menetapkan semaksimum lima aksara panjang akhiran perkataan. Jadual 10 menunjukkan prestasi penandaan mengguna MMT trigram dengan peramal GK perkataan anu mengguna huruf akhiran perkataan dengan maksimum panjang akhiran yang pelbagai. Sebagai perbandingan, ramalan perkataan anu melalui awalan perkataan mengatasi prestasi ramalan melalui akhiran perkataan. Awalan perkataan dalam Bahasa Melayu banyak terdiri daripada pengimbuhan yang dominan seperti dalam Jadual 3, iaitu imbuhan awalan *peN-*, *pe-*, *peR-*, dan *ke-* sentiasa memberi kategori kata nama. Begitu juga dengan imbuhan awalan *meN-*, *beR-*, *di-*, *mempeR-*, dan *dipeR-* sentiasa memberi kategori kata kerja. Satu-satunya awalan yang boleh memberi dua kategori adalah *teR-*, iaitu sama ada kata kerja atau kata adjektif. Begitu juga dengan apitan *ke-...-an* memberikan tanda GK sama ada kata nama atau kata adjektif.

JADUAL 9. Prestasi penandaan GK oleh MMT trigram dengan ramalan perkataan anu melalui awalan perkataan

Panjang awalan perkataan	Kejituan keseluruhan	Kejituan GK untuk perkataan diketahui	Kejituan GK untuk perkataan anu
1	93.1	98.6	62.2
2	93.8	98.6	66.9
3	94.0	98.6	67.9
4	93.9	98.6	67.6
5	93.9	98.6	67.1
6	93.8	98.6	66.9
7	93.8	98.6	66.8
8	93.8	98.6	66.9
9	93.8	98.6	66.9
10	93.8	98.6	66.9

JADUAL 10. Kejituan penandaan GK oleh MMT trigram dengan ramalan perkataan anu melalui akhiran perkataan

Panjang akhiran perkataan	Kejituan keseluruhan	Kejituan GK untuk perkataan diketahui	Kejituan GK untuk perkataan anu
1	92.1	98.6	55.4
2	92.1	98.6	55.4
3	92.5	98.5	58.7
4	92.5	98.5	58.9
5	92.7	98.5	60.0
6	92.7	98.5	59.9
7	92.7	98.5	59.9
8	92.7	98.5	59.9
9	92.7	98.5	59.9
10	92.7	98.5	59.9

Prestasi terbaik untuk meramal GK perkataan anu melalui gabungan adalah semaksimum tiga aksara awalan dan juga tiga aksara akhiran. Ini menunjukkan prestasi meramal GK perkataan anu melalui gabungan dilihat cenderung dipengaruhi oleh awalan perkataan. Ini kerana ramalan melalui awalan perkataan dengan panjang maksimum tiga aksara mengatasi ramalan melalui akhiran perkataan. Bahkan, ramalan melalui awalan perkataan mempunyai prestasi yang lebih sedikit berbanding dengan prestasi ramalan mengguna gabungan awalan dan akhiran perkataan. Jadual 11 menunjukkan prestasi penandaan mengguna MMT trigram dengan peramal untuk GK perkataan anu mengguna-gabungan awalan dan akhiran perkataan dengan kombinasi maksimum panjang yang sama.

JADUAL 11. Kejituan penandaan GK oleh MMT trigram dengan ramalan perkataan anu melalui apitan perkataan

Panjang awalan perkataan	Panjang akhiran perkataan	Kejituan keseluruhan	Kejituan GK untuk perkataan diketahui	Kejituan GK untuk perkataan anu
1	1	92.7	98.6	60.1
2	2	93.5	98.6	64.9
3	3	93.7	98.6	66.7
4	4	93.7	98.6	66.2
5	5	93.7	98.6	65.9
6	6	93.6	98.6	65.7
7	7	93.6	98.6	65.7
8	8	93.6	98.6	65.8
9	9	93.6	98.6	65.8
10	10	93.6	98.6	65.8

Melihat kepada panjang maksimum tiga aksara sebagai yang terbaik untuk ramalan mengguna awalan perkataan dan semaksimum lima bagi ramalan mengguna akhiran perkataan, maka gabungan angka tiga dan lima ini diuji dalam ramalan mengguna kaedah gabungan. Keputusannya adalah 66.8% kejituan yang menunjukkan dapatan yang bertentangan dengan apa yang dirasai secara intuitif bagi menganggap gabungan melakukan yang baik.

KESIMPULAN

Alat penanda GK berpenyelia yang sedia ada diuji ke atas Bahasa Melayu. Tiga alatan penanda tersebut diguna iaitu *TnT*, *MaxEnt* dan *SVMTool*. Anotasi korpus Bahasa Melayu bagi domain kesihatan dilakukan kerana ketiga-tiga alatan tersebut memerlu korpus bertanda bertujuan untuk latihan dan ujian bagi mendapat tanda aras atau perbandingan prestasi. Prestasi *SVMTool* mengatasi *TnT* dan *MaxEnt* bagi kejituan keseluruhan serta kejituan penandaan perkataan anu. Keupayaan *MaxEnt* pula mengatasi *TnT* bagi kejituan keseluruhan serta kejituan penandaan perkataan anu. Oleh itu penandaan GK Bahasa Melayu bagi domain spesifik ketika ini adalah dicadangkan mengguna alatan sedia ada melalui penyediaan korpus bertanda Bahasa Melayu.

RUJUKAN

Abdullah, H., S. Rohani, L. J., Ayob, R. dan Osman, Z. 2006. *Sintaksis siri pengajaran dan pembelajaran Bahasa Melayu*. Kuala Lumpur: PTS Professional.

Abney, S., Schapire, R. E. and Singer, Y. 1999. Boosting applied to tagging and PP attachment. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and very Large Corpora.* New Jersey: Association for Computational Linguistics, 38-45

Arbak, O. 2005. *Kamus komprehensif Bahasa Melayu*. Shah Alam: Oxford Fajar.

Brants, T. 2000. TnT: a statistical part-of-speech tagger. *Proceedings of the 6th conference on Applied natural language processing*. Pennsylvania: Association for Computational Linguistics, 224-231.

Brill, E. 1995. Transformation-based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4): 543-566.

Brill, E. and Wu, J. 1998. Classifier combination for improved lexical disambiguation. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. Pennsylvania: Association for Computational Linguistics, 191-195.

Collins, M. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. *Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing - Volume 10*. Pennsylvania: Association for Computational Linguistics, 1-8.

Daelemans, W., Zavrel, J., Berck, P. and Gillis, S. 1996. MBT: A Memory-Based Part of Speech Tagger-Generator. *Proceedings of the 4th Workshop on very Large Corpora*. Copenhagen: University of Copenhagen, 14-27.

Giménez, J. and Màrquez, L. 2003. Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. *Proceeding of the 4th Recent Advances in Natural Language Processing*. Borovets: Bulgarian Academy of Sciences, 153-163

Giménez, J. and Màrquez, L. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon: European Language Resources Association, 43-46.

Halteran van, H., Zavrel, J. and Daelmans, W. 1998. Improving Data Driven Wordclass Tagging by System Combination. *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. Pennsylvania: Association for Computational Linguistics, 491-497.

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

Hock, O. Y. 2009. *Kamus Dwibahasa*. Petaling Jaya: Pearson Longman.

Jurafsky, D. and Martin, J. H. 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 2nd ed. New Jersey: Prentice Hall.

Knowles, G. and Zuraida. M. D. 2006. *World Class in Malay: A Corpus-based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Lee, S., Tsujii, J. and Rim, H. 2000. Part-of-Speech Tagging Based on Hidden Markov Assuming Joint Independence. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Pennsylvania: Association for Computational Linguistics, 263-269.

Manning, C. D. and Schutze, H. 1999. *Foundations of statistical natural language processing*. Massachusetts: MIT Press Cambridge.

Màrquez, L. and Rodríguez, H. 1997. Automatically Acquiring a Language Model for POS Tagging Using Decision Trees. In: Nicolov, Nicolas and Ruslan Mitkov (ed.) *Recent Advances in Natural Language Processing: Volume II: Selected papers from RANLP '97,* 31-39. Amsterdam: John Benjamins Publishing Company.

Màrquez, L., Rodríguez, H., Carmona, J. and Montolio, J. 1999. Improving POS Tagging Using Machine-Leraning Techniques. *Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing and very Large Corpora*. Maryland: SIGDAT, 53-62.

Nakagawa, T., Kudoh, T. and Matsumoto, Y. 2001. Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines. *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*. Tokyo: AFNLP, 325-331.

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

Ratnaparkhi, A. 1996. A Maximum Entropy Part-Of-Speech Tagger. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Pennsylvania: University of Pennsylvania, 133-142.

Weischedel, R., Schwartz, R., Palmucci, J., Meteer, M. and Ramshaw, L. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistic,* 19(2): 361-382.

Samuelsson, C. 1996. Handling sparse data by successive abstraction. *Proceedings of the 16th International conference on Computational linguistics - Volume 2*. Copenhagen: COLING, 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.

Ranaivo-Malançon, B. 2008. Issues in building a Malay part of speech tag-set. *Proceeding of the 2nd International MALINDO Workshop*. Cyberjaya: Multimedia University, 104-108.

Toutanova, K. and Manning, C. D. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and very Large Corpora*. Pennsylvania: Association for Computational Linguistics, 63-70.

Thede, S. M. and Harper, M. P. 1999. A second-order Hidden Markov Model for part-of-speech tagging. *Proceedings of the Conference for 37th Annual Meeting of the Association for Computational Linguistics.* Pennsylvania: Association for Computational Linguistics, 175-182.

Hassan Mohamed,
Assoc. Prof. Dr. Nazlia Omar,
Assoc. Prof. Dr. Mohd. Juzaiddin Ab. Aziz
Knowledge Technology Research Group,
Centre for Artifial Intelligence and Technology (CAIT),
Faculty of Information Science & Technology,
Universiti Kebangsaan Malaysia
hassan.dbangi@yahoo.com, nazlia@ukm.edu.my, juzaiddin@ukm.edu.my