http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 6 No. 2, December 2017:
e-ISSN: 2289-2192

MALAY PART OF SPEECH TAGGING USING RULED-BASED APPROACH

NUR ASHIKIN HALID
NAZLIA OMAR

ABSTRACT

The research on part of speech (POS) tagging has been widely applied and used through a variety of approaches, particularly for European languages. But it is more challenging for Asian languages, especially Malay as it has some element of modification from other languages such as English and Arabic. Among the issues that often occur in POS tagging are the existence of ambiguous words and unknown words. Meanwhile, the lack of rules in the existing work has become a major problem in Malay POS tagging. Therefore, this research aims to develop new rules for Malay POS tagging and to compare the performance of this new development with the existing gold standard. This process begins with the collection and selection of the corpus using secondary data, obtained from online daily news which covers several domains. Next, the corpus has gone through the process of pre-processing in raw text of article form which include sentence splitter and tokenization process to generate an unlabeled corpus. POS tag dictionary also has been constructed to form a lexicon that only consists of root words. The rule development process involves detailing every type of POS tag to its suitable rules and get the best rules ordering for each type of this POS. A total of 30 rules including affixation rules and 16 word type relations have been developed in this process. The evaluation process is used to test the precision of the developed POS tagger and to get the best rules ordering. The POS tagging result is compared with existing gold standard. Overall, the test showed good result with an accuracy of 93.06% compared to the gold standard performance of 77.17%. Hence, this research showed better accuracy compared with the gold standard and at the same time, it proves that the addition of a new rules and rules ordering among the factors that contributed to the higher precision in tagging Malay corpus. As an improvement in future studies, the use of compound words should be taken into account because most of these words are used in most news sources. In addition, corpus from social media sources can be used because the content of information disseminated through social media is fast and up-to-date even though the language used for this resource is mostly informal and confronts with noise data issues.

Keywords: Malay Part of Speech Tagging, Ruled-Based Approach, Part of Speech Rule, Word Type Relations Rule.

PENANDAAN GOLONGAN KATA BAHASA MELAYU MENGGUNAKAN PENDEKATAN BERASASKAN PETUA

ABSTRAK

Kajian mengenai penandaan Golongan Kata (GK) telah banyak dilakukan secara meluas dan digunakan melalui pelbagai pendekatan terutama bagi bahasa Eropah. Namun ianya lebih mencabar bagi bahasa Asia khususnya bahasa Melayu kerana beberapa perkataan Melayu mempunyai unsur pengubahsuaian daripada bahasa lain seperti bahasa Inggeris dan bahasa Arab. Isu yang sering timbul dalam proses penandaan GK bahasa Melayu adalah kewujudan perkataan kabur dan perkataan yang tidak diketahui. Selain itu, masalah utama dalam penandaan GK bahasa Melayu adalah kekurangan petua dalam kajian sedia ada. Justeru, kajian ini bertujuan membangunkan petua baru bagi penandaan GK Bahasa Melayu dan membandingkan prestasi penandaan GK bahasa Melayu berasaskan petua dengan piawaian emas sedia ada. Proses ini bermula dengan pengumpulan dan pemilihan korpus menggunakan data sekunder yang diperoleh daripada Berita Harian secara atas talian meliputi pelbagai domain. Korpus seterusnya melalui pra-pemprosesan di mana artikel dalam bentuk teks mentah melalui proses pemisahan ayat dan pentokenan supaya korpus tidak bertanda dapat dihasilkan. Kamus GK juga telah dibina bagi membentuk leksikon yang hanya terdiri daripada kata akar sahaja. Proses pembangunan petua pula merupakan proses memperincikan setiap jenis GK kepada petuanya yang tersendiri dan memberi susunan aturan kedudukan kepada

setiap jenis GK ini. Sebanyak 30 petua GK termasuk petua imbuhan dan 16 petua hubungan kata dibangunkan dalam proses ini. Proses penilaian pula dilaksanakan bagi melihat ketepatan petua GK yang dibangunkan dan aturan susunan petua GK yang terbaik serta membuat perbandingan hasil penandaan GK dengan piawaian emas sedia ada. Secara keseluruhannya, pengujian ini memberikan hasil yang baik dengan nilai ketepatan 93.06% berbanding prestasi piawaian emas iaitu 77.17%. Oleh yang demikian, kajian ini menunjukkan ketepatan yang lebih baik berbanding dengan piawaian emas dan pada masa yang sama membukti bahawa penambahan petua baru serta susun atur kedudukan petua merupakan antara faktor yang menyumbang kepada nilai ketepatan yang lebih tinggi dalam penandaan GK bahasa Melayu. Sebagai penambahbaikan dalam kajian masa hadapan, penggunaan perkataan majmuk perlu diambilkira kerana kebanyakan perkataan ini digunakan dalam kebanyakan sumber berita. Selain itu juga, korpus daripada sumber sumber media sosial boleh digunakan kerana walaupun bahasa yang digunakan bagi sumber ini kebanyakannya tidak formal dan berdepan dengan isu data hingar namun kandungan maklumat yang disebarkan melalui media sosial ini adalah cepat dan terkini.

Kata Kunci: Penandaan Golongan Kata Bahasa Melayu, Pendekatan Petua, Petua Golongan Kata, Petua Hubungan Kata.

PENGENALAN

Kajian mengenai penandaan Golongan Kata (GK) telah banyak dilakukan secara meluas dan digunakan melalui pelbagai pendekatan terutama bagi bahasa Eropah. Namun ianya lebih mencabar bagi bahasa Asia khususnya bahasa Melayu kerana beberapa perkataan Melayu mempunyai unsur pengubahsuaian daripada bahasa lain seperti bahasa Inggeris dan bahasa Arab. Isu yang sering timbul dalam proses penandaan GK bahasa Melayu adalah kewujudan perkataan kabur dan perkataan yang tidak diketahui. Selain itu, masalah utama dalam penandaan GK bahasa Melayu adalah kekurangan petua dalam kajian sedia ada. Justeru, kajian ini bertujuan membangunkan petua baru bagi penandaan GK Bahasa Melayu dan membandingkan prestasi penandaan GK bahasa Melayu berasaskan petua dengan piawaian emas sedia ada.

Kajian ini merupakan salah satu daripada kajian yang menjalankan proses menganotasi atau memberi tanda nama dalam ayat untuk setiap kelas token atau perkataan seperti kata nama, kata kerja, kata sifat (adjektif) dan kata keterangan bergantung kepada hubungan perkataan dan juga definisi ayat (Alfred et al., 2013). Kajian ini memberi fokus kepada morfologi dan sintaksis dalam bidang linguistik Melayu memandangkan kedua-dua kriteria tersebut bertepatan dengan kehendak kajian yang dijalankan ini. Morfologi merupakan kajian mengenai cara perkataan dibina daripada unit-unit kecil yang dipanggil morfem (Jurafsky & Martin, 2011); morfem pula terbahagi kepada dua kelas utama iaitu akar dan imbuhan. Sintaksis pula merupakan istilah bagi cara susunan dan urutan dalam ayat. Ianya memerlukan tatabahasa dan penghurai (bagi menghuraikan perkataan dan frasa kepada beberapa bahagian untuk memahami maksud perkataan dan hubungan antara perkataan dalam ayat) yang memberi tumpuan kepada analisis perkataan dalam ayat bagi mempamerkan struktur tatabahasa dalam ayat tersebut (Preeti & Sidhu, 2013).

Menurut Kumawat dan Jain (2015), pendekatan berasaskan petua menggunakan satu set peraturan yang bertulis untuk digunakan sebagai penanda GK bagi perkataan berdasarkan peraturan yang telah disediakan. Dalam kajian ini, sebanyak 30 petua GK dan 16 petua hubungan kata dibina menggunakan set GK bahasa Melayu. Manakala proses pembentukan kata menggunakan pengimbuhan yang terdiri daripada imbuhan awalan, akhiran, apitan dan sisipan yang akan menghasilkan suatu bentuk kata terbitan (Nik Safiah et al., 2015).

LATAR BELAKANG

Kajian berkaitan dengan penandaan GK bagi bahasa Melayu dalam kajian ini mengambilkira penggunaan penandaan GK bagi bahasa Melayu dan bahasa Indonesia memandangkan kedua-dua bahasa tersebut berasal dari rumpun keluarga bahasa Austronesian yang sama (Alfred et

al., 2013) dan mempunyai ciri-ciri bahasa dan tatabahasa yang hampir serupa. Kajian ini menggunakan pendekatan berasaskan petua bagi kajian yang telah dilaksanakan oleh Alfred et al. (2013) dan dijadikan piawaian emas bagi membandingkan prestasi penandaan GK yang dibina. Kajian yang menggunakan pendekatan berasaskan petua dipilih kerana masih terdapat kekurangan petua dalam petua sedia ada sama ada petua yang melibatkan penambahan dalam kategori GK atau petua hubungan perkataan. Selain itu, antara masalah lain ialah yang melibatkan imbuhan sisipan yang tidak dinyatakan dalam kebanyakan pendekatan sedia ada kerana ia dianggap sebagai imbuhan yang tidak popular dan tidak produktif (Abdullah, 2006). Jadual 1 menunjukkan ringkasan perbandingan kajian-kajian yang telah dilakukan dalam penandaan GK bahasa Melayu beserta kekuatan dan kelemahan kajian masing-masing.

JADUAL 1. Perbandingan kajian-kajian lepas

Penyelidik	**Kaedah Kajian**	**Kekuatan**	**Kelemahan**	**Hasil/ Peratusan Kejituan**
Alfred et al. (2013)	RPOS (berasaskan petua)	▪ Dapat menanda perkataan anu dengan baik pada kejituan yang tinggi. ▪ Hasil keputusan kejituan adalah tinggi berbanding dengan kaedah statistik.	▪ Memerlukan usaha dan tenaga yang banyak bagi menandakan korpus secara manual. ▪ Hasil keputusan kejituan yang rendah bagi perkataan pinjaman daripada bahasa Inggeris. ▪ Tidak mempertimbangkan imbuhan sisipan.	▪ 89% bagi artikel Melayu dan 86% bagi artikel bio-perubatan.
Zuraidah (2010)	MALEX (berasaskan kaedah penormalan (*normalizing*), menghuraikan (*parsing*) dan memberi kata akar (*stemming*) berdasarkan pendekatan sintaksis dan didorong oleh data (*data-driven*)	▪ Menggunakan beberapa jenis sumber data (novel, suratkhabar, teks ucapan Perdana Menteri dan teks akademik) dan bilangan data yang banyak. ▪ Mengambilkira keperluan tatabahasa dan sintaksis dalam pembinaan MALEX.	▪ Memerlukan usaha dan tenaga yang banyak bagi mencari dan memilih data yang sesuai dengan bilangan yang agak banyak.	▪ Tiada peratus kerana kajian ini bertujuan membina pangkalan data bagi analisis teks melayu. Sebanyak 120,000 perkataan telah ditandakan dalam kajian ini.
Mohd Pouzi dan Syarifah Fatem Na'imah (2014)	Pendekatan Berasaskan Petua (Morfologi)	▪ Keputusan penandaan GK yang baik menggunakan aspek morfologi.	▪ Data (laporan polis) yang digunakan tidak menggunakan bahasa formal ▪ Penandaan GK hanya melibatkan empat penanda GK sahaja iaitu kata nama, kata kerja, kata sifat dan kata adverba. ▪ Perkataan yang mempunyai unsur imbuhan dan kependekan tidak dapat ditandakan dengan betul.	▪ 88.4% bagi laporan polis.
Juhaida et al.	Pendekatan	▪ Penggunaan teknik	▪ Analisis hanya	▪ 92.86%

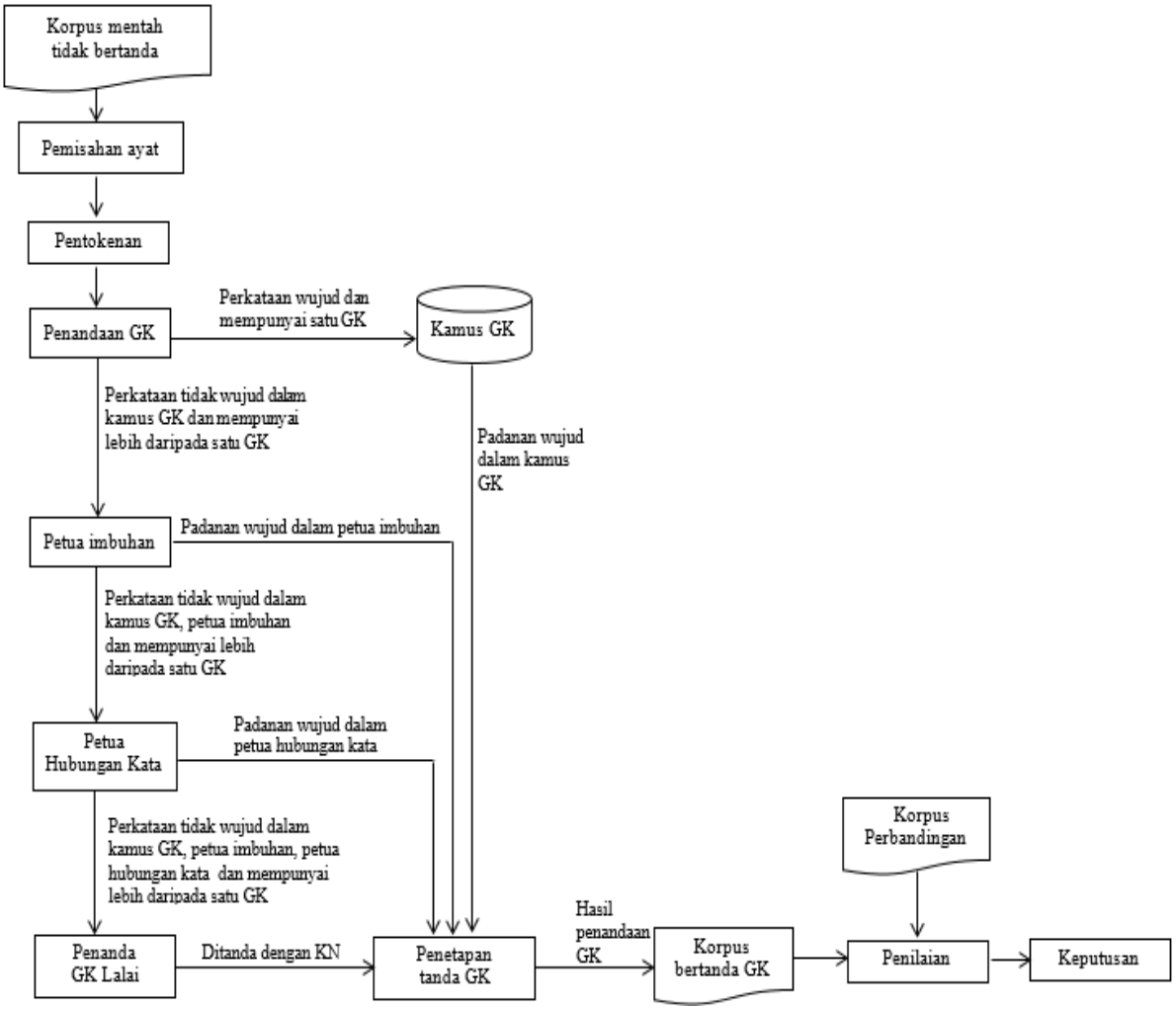
(2013)	Berasaskan Petua (analisis morfologi pada algoritma pembelajaran mesin iaitu kaedah Pepohon Keputusan dan Jiran Terdekat (*Nearest Neighbor*)	yang fleksibel (kaya dengan ciri-ciri perwakilan) bagi Pepohon Keputusan. ▪ Mampu mengenalpasti data yang tidak ditandakan dalam data ujian dengan menggunakan unsur morfologi	mengambilkira aspek morfologi sahaja tanpa unsur lain seperti semantik.	melalui kaedah Pepohon Keputusan
Mohamed dan Rohana (2015)	Membina Kamus Penandaan Golongan Kata Bahasa Melayu Menggunakan Bahasa *WordNet* dan Korpus Bahasa Indonesia	▪ Menggunakan sumber data daripada pelbagai jenis dengan jumlah yang banyak. ▪ Gabungan penggunaan *WordNet* dan korpus bahasa Indonesia saling melengkapi dan menghasilkan keputusan penandaan GK yang tinggi	▪ Penggunaan sumber data yang melibatkan tahap kualiti sederhana. ▪ Penggunaan penandaan GK yang agak terhad daripada WordNet dan korpus bahasa Indonesia.	▪ Tiada peratus kerana kajian ini bertujuan menghasilkan data sebagai input kepada analisis teks melayu. Sebanyak 25,778 perkataan telah ditandakan dalam kajian ini.
Rashel et al. (2014)	Pendekatan Berasaskan Petua	▪ Menggunakan sumber data daripada pelbagai jenis dengan jumlah yang banyak. ▪ Menggunakan lima peringkat teknik (pentokenan, pengecaman entiti nama, penandaan GK bagi perkataan tertutup dan terbuka, peraturan nyahkabur dan penyelesai (*resolver*)) menghasilkan keputusan penandaan GK yang tinggi.	▪ Memerlukan usaha dan tenaga yang banyak bagi menandakan korpus secara manual. ▪	▪ 79% melalui penggunaan pengecaman entiti nama.
Widhiyantil dan Agus (2012)	Penandaan Golongan Kata Bahasa Indonesia Menggunakan Model Markov Tersembunyi dan Pendekatan Berasaskan Petua	▪ Gabungan penggunaan kedua-dua pendekatan berjaya menghasilkan keputusan penandaan GK yang tinggi dalam korpus yang sama.	▪ Penandaan GK tidak dapat menandakan perkataan anu bagi teks yang tidak ada dalam korpus yang sama. ▪ Penggunaan jenis GK yang tidak banyak.	▪ 92.2% bagi jenis korpus yang berbeza.

Kajian bagi piawaian emas telah menghasilkan banyak petua bagi penandaan GK bahasa Melayu namun jumlah penghasilan petua masih tidak mencukupi dan tidak menyeluruh seperti tidak mengambilkira imbuhan sisipan, yang walaupun jarang digunakan tetapi masih boleh ditemui dalam sesetengah teks bahasa Melayu. Kebanyakan kajian juga menumpukan kepada korpus dalam satu domain sahaja (Alfred et al., 2013; Mohd Pouzi dan Syarifah Fatem Na'imah, 2014) seperti domain perubatan, kesihatan dan laporan polis. Ini menyebabkan hasil

penandaan GK tidak dapat dilihat secara menyeluruh kerana keputusan penandaan GK terpakai kepada domain tertentu sahaja. Terdapat juga sebahagian kajian yang hanya menumpukan kepada satu kriteria dalam mengkategorikan perkataan sahaja (Juhaida et al., 2013; Mohd Pouzi dan Syarifah Fatem Na'imah, 2014; Zuraidah, 2010) iaitu sama ada kajian tersebut bertumpu kepada unsur morfologi atau unsur sintaksis sahaja. Justeru, kajian ini mengambilkira penambahan jumlah petua dan penggunaan korpus dari pelbagai domain agar penandaan GK dapat dilihat dari segi pelbagai variasi. Secara puratanya, setiap korpus mempunyai kira-kira 200 token termasuk tanda bacaan dan simbol. Korpus-korpus yang dikumpulkan hanya memetic sebahagian daripada artikel-artikel daripada sumber berita untuk tujuan kajian ini sahaja. Penambahan jenis GK dan perkataan dalam leksikon juga turut dititikberatkan selain memasukkan unsur morfologi dan sintaksis dalam kajian ini.

METOD KAJIAN

Metod kajian adalah penting bagi memberikan gambaran yang lebih jelas dan secara menyeluruh tentang sesuatu kaedah bagi mendapatkan maklumat yang digunakan bagi mencapai objektif kajian (Haroon, 2010). Secara ringkasnya, proses tersebut bermula dengan penyediaan korpus yang meliputi korpus latihan yang akan digunakan dalam pembangunan petua dan korpus ujian yang akan digunakan dalam proses penilaian. Seterusnya, proses pra-pemprosesan diikuti dengan pembangunan petua yang menjadi tulang belakang dalam kajian ini dan akhir sekali penilaian. Rangka kerja aliran proses kajian adalah seperti yang ditunjuk dalam Rajah 1.



RAJAH 1. Rangka kerja aliran proses kajian

Berdasarkan Rajah 1, korpus mentah tidak bertanda disediakan sebelum proses penandaan GK dilaksanakan. Seterusnya, proses pra-pemprosesan korpus yang melibatkan proses pemisahan ayat dan pentokenan dilakukan. Dalam kajian ini, kamus penandaan GK ini telah dibina secara manual yang hanya terdiri daripada senarai perkataan atau token yang mempunyai kata akar sahaja yang dengan penanda GK masing-masing bagi membentuk leksikon. Penandaan GK ke atas korpus tidak bertanda dimulakan dengan memadankan perkataan dengan kamus GK dahulu dan jika perkataan tersebut wujud dan mempunyai satu penanda GK, maka penandaan GK ditetapkan kepada perkataan tersebut.Sekiranya perkataan tersebut tidak wujud dalam kamus dan mempunyai lebih daripada satu penanda GK, perkataan dipadankan mengikut turutan petua imbuhan. Jika perkataan tersebut tidak wujud dalam kamus, mempunyai lebih daripada satu penanda GK dan masih tidak berjaya ditanda mengikut petua imbuhan, perkataan dipadankan mengikut senarai hubungan kata yang ditetapkan. Jika tiada penanda GK yang sesuai berdasarkan hubungan kata yang diberikan, proses penandaan GK akan diteruskan dengan menanda perkataan yang tidak dapat dikenalpasti kepada penanda GK lalai iaitu kata nama.

Hasil penandaan GK dipaparkan melalui penghasilan korpus bertanda dengan setiap perkataan ditandakan dengan jenis GK masing-masing. Proses penilaian kemudian dijalankan dengan membandingkan hasil penandaan GK tersebut dengan korpus perbandingan yang telah ditandakan secara manual terlebih dahulu. Keputusan akhir penilaian merupakan hasil perbandingan prestasi penandaan GK dengan korpus perbandingan.Proses pra-pemprosesan korpus, pembangunan petua dan susunan aturan petua GK yang terlibat dalam kajian ini diterangkan secara terperinci selepas ini.

1. Pra-pemprosesan

Seperti yang diterangkan sebelum ini, korpus yang dipilih akan melalui proses pemisahan ayat dan pentokenan. Dalam proses ini, korpus yang terdiri daripada teks mentah akan dipecahkan kepada bentuk ayat dan pemisahan kepada ayat demi ayat ini dilakukan dengan mengenalpasti tanda noktah, tanda seru atau tanda soal yang terdapat di setiap penghujung perkataan dalam sesuatu ayat. Tanda noktah, tanda seru dan tanda soal berfungsi sebagai sempadan yang memisahkan antara satu ayat dengan ayat yang lain dalam satu teks. Manakala dalam proses pentokenan, setiap perkataan, tanda bacaan dan simbol dipisahkan melalui ruang kosong (*white space*) dan ruang kosong ini juga berfungsi sebagai sempadan yang memisahkan antara satu perkataan dengan perkataan atau simbol yang lain dalam satu teks.

2. Pembangunan petua

Proses pembangunan petua adalah proses terpenting dalam kajian ini kerana pembangunan petua merupakan proses utama dalam memastikan objektif kajian dapat dicapai. Kajian ini menggunakan semula 15 jenis kelas kata dan 14 petua hubungan kata sedia ada yang telah dibina oleh Alfred et al. (2013) dalam penandaan GK piawaian emas dan sebanyak 15 petua baru GK dan dua petua hubungan kata baru telah ditambah dalam kajian ini seperti yang ditunjukkan dalam Jadual 2.

JADUAL 2. Senarai petua GK dan hubungan kata

Bil.	Petua GK		Bil.	Hubungan Kata	
1.	Kata Nama (KN)		1.	Kata Nama (KN)	
2.	Kata Sifat (KA)		2.	Kata Sifat (KA)	
3.	Kata Kerja (KK)		3.	Kata Kerja (KK)	
4.	Kata Penegas (KPN)		4.	Kata Penegas (KPN)	
5.	Kata Pembenda (KPB)		5.	Kata Pembenda (KPB)	
6.	Kata Penekan (KPT)		6.	Kata Adverba (KAD)	
7.	Kata Adverba (KAD)		7.	Kata Arah (KAR)	
8.	Kata Arah (KAR)	Jenis kelas kata sedia ada dari Alfred et al. (2013)	8.	Kata Sendi (KS)	Petua hubungan kata sedia ada dari Alfred et al. (2013)
9.	Kata Sendi (KS)		9.	Kata Bantu Aspek (KBA)	
10.	Kata Bantu Aspek (KBA)		10.	Kata Bilangan (KBIL)	
11.	Kata Bilangan (KBIL)		11.	Kata Hubung (KH)	
12.	Kata Hubung (KH)		12.	Kata Penguat (KP)	
13.	Kata Penguat (KP)		13.	Kata Tanya (KTY)	
14.	Kata Tanya (KTY)		14.	Penanda Wacana (PW)	
15.	Penanda Wacana (PW)		15.	Kata Nafi (KNF)	Petua hubungan kata yang ditambah
16.	Kata Penentu (TEN)		16.	Kata Pemeri (KPM)	
17.	Kewujudan (ADA)				
18.	Kata Pinjaman (PIN)				
19.	Kata Bantu Ragam (KBR)				
20.	Kata Nama Khas (KNK)				
21.	Kata Ganti Nama (KGN)				
22.	Ganti Nama Relatif (GNR)				
23.	Gelaran (GEL)	Petua baru yang ditambah			
24.	Kependekan (KEP)				
25.	Kata Nafi (KNF)				
26.	Kata Pembenar (KPR)				
27.	Penjodoh Bilangan (JDH)				
28.	Kata Pemeri (KPM)				
29.	Kata Perintah (KTP)				
30.	Simbol (SYM)				

Jumlah petua yang dibina bagi GK adalah sebanyak 30 manakala petua bagi hubungan kata pula adalah 16. Walau bagaimanapun, hanya lima petua GK sahaja yang dipertimbangkan bagi petua yang mempunyai imbuhan awalan, apitan, akhiran dan sisipan iaitu kata nama (KN), kata kerja (KK), kata adjektif atau sifat (KA), kata penekan (KPT) dan kata pembenda (KPB). Ini kerana dalam pemilihan perkataan-perkataan yang mempunyai kata akar, hanya perkataan yang terdiri daripada GK KN, KK dan KA sahaja yang dipertimbangkan bagi membezakan perkataan akar ini dengan perkataan yang mempunyai imbuhan awalan, apitan, akhiran dan sisipan. Manakala bagi GK KPT dan KPB yang melibatkan penggunaan akhiran *–nya*, penekanan hanya diberikan kepada syarat penandaan imbuhan bagi GK KN sahaja. Ini bermaksud, jika sesuatu perkataan terlibat dengan salah satu syarat bagi petua imbuhan GK KN dan berakhir dengan *–nya*, maka perkataan tersebut akan terus ditandakan dengan GK KPT. Sebaliknya jika sesuatu perkataan itu tidak terlibat dengan petua imbuhan GK KN tetapi berakhir dengan *–nya*, maka perkataan tersebut akan terus ditandakan dengan GK KPB. Ketiga-tiga GK KN, KK dan KA ini mempunyai Imbuhan sisipan merupakan petua baru yang ditambah dalam ketiga-tiga jenis GK KN, KK dan KA. Jadual 3 menunjukkan contoh perincian petua imbuhan yang melibatkan imbuhan awalan, akhiran, sisipan dan apitan bagi KN dan Jadual 4 pula menunjukkan contoh perincian petua hubungan kata yang digunakan dalam kajian ini.

JADUAL 3. Petua imbuhan Kata Nama

Peraturan	Imbuhan (Awalan)	Karakter seterusnya	Urutan Karakter	Imbuhan (Akhiran) Boleh berakhir dengan	
1a	pe	ny, ng, r, l dan w	a-z	an	-
1b	pem	b dan p	a-z	an	-
1c	pen	d,c,j, sy dan z	a-z	an	-
1d	peng	g, kh, h,k, dan vokal	a-z	an	-
1e	penge	-	a-z (3 ke 4 karakter)	an	-
1f	pel atau ke	-	a-z	an	-
1g	juru, maha, tata, pra, swa, tuna, eka, dwi, tri, panca, pasca,pro,anti,poli, auto, sub,supra, mono	-	a-z	-	-
1h	tidak bermula dengan me, meng, mem, menge, ber, be, di, diper	-	a-z	-	an, at, in,wan, wati, isme, isasi, logi, tas, man, nita, ik, is, al
1i	gel, tel,	ang,ap,un	a-z	-	-
1j	kel	ang,op, eng,ab	a-z	-	-
1k	gem, tem, sem,	al, un, an, in	a-z	-	-
1l	kem	un	bukan a,i,o,u	-	-
1m	ger, ker, ser, cer	ig, ab, ul, up, ic	a-z	-	-
1n	baha, saha	gi, ja	-	an	-

JADUAL 4. Contoh petua hubungan kata

Jenis GK	Petua	Contoh penggunaan
Kata Nama	Jika token bermula dengan huruf besar atau kecil, dan jenis kelas kata selepas *Kata Nama* terdiri daripada jenis kelas kata *Kata Adjektif, Kata Adverba, Kata Kerja, Kata Nama, Kata Sendi* atau *Kata Ganti Nama*, maka jenis GK adalah Kata Nama	**Kali\KN terakhir\KA** Bank Negara menurunkan kadar campur tangan **Pasaran\KN menjangkakan\KK** BankNegara membuat pengumuman itu pada hari ini..
Kata Kerja	Jika token bermula dengan huruf besar atau kecil, dan jenis kelas kata selepas *Kata Kerja* terdiri daripada jenis kelas kata *Kata Bantu Ragam, Kata Adverba, Kata Nama, Kata Penekan, Kata Pembenda, Kata Adjektif* atau *Kata Ganti Nama*, maka jenis GK adalah Kata Kerja	Perak **mengenalpasti\KK lebih\KAD** 800 rakyat miskin
Kata Adjektif	Jika token bermula dengan huruf besar atau kecil, dan jenis kelas kata selepas *Kata Adjektif* terdiri daripada jenis kelas kata *Kata Penguat* atau *Kata Sendi*, maka jenis GK adalah Kata Adjektif	Khabar angin tersebar **luas\KA mengenai\KS** kemungkinan
Kata Penekan	Jika token bermula dengan huruf besar atau kecil, dan jenis kelas kata selepas *Kata Penekan* terdiri daripada jenis kelas kata *Kata Adverba*, *Kata Nama* atau *Kata Hubung*, maka jenis GK adalah Kata Penekan	Beliau melihat sendiri penumpang yang **kebanyakannya\KPT kanak-kanak\KN**

Walaupun KPT dan KPB tidak terlibat secara langsung dalam proses pengimbuhan ini, tetapi kedua-dua jenis GK ini mempunyai kaitan dengan KN, KK dan KA. Ini kerana, KPT bergantung kepada penandaan perkataan bagi GK KN dan KPB pula bergantung kepada penandaan perkataan bagi GK selain daripada KN. Cara penandaan kajian ini adalah sedikit berbeza dengan cara penandaan GK yang dijalankan oleh Alfred et al. (2013). Algoritma penandaan GK bagi kajian ini adalah seperti Jadual 5:

JADUAL 5. Algoritma bagi penandaan GK

Algoritma bagi penandaan GK

Input : *string* a *(tidak bertanda)*
Hasil : *string* a *dan* b*(yang bertanda)*

IF (token is KATA_AKAR and ADA_DALAM_KAMUS) **THEN**
  tag is PADANAN_DALAM_KAMUS
ELSE IF (token is KATA_IMBUHAN) **THEN**
  tag is KN or KK or KA or KPT or KPB
ELSE IF (token is KATA_BUKAN_IMBUHAN) **THEN**
  tag is SELAIN_KN_or_KK_or_KA_or_KPT_or_KPB
ELSE IF (token is KATA_BUKAN_IMBUHAN) **THEN**
  tag is PADANAN_HUBUNGAN_KATA
ELSE
  tag is KN

3. Susunan aturan petua GK

Susunan aturan petua bagi 30 jenis GK ini amat penting dalam menentukan kejituan penandaan GK. Penyusunan dan aturan kedudukan petua yang kurang tepat akan mengurangkan peratusan kejituan penandaan GK terhadap sesuatu perkataan. Penyusunan petua juga tidak dinyatakan dalam kajian yang dilakukakn oleh Alfred et al. (2013). Dalam kajian ini, pengujian prestasi penandaan GK turut dilakukan bagi menilai prestasi kejituan penandaan GK terhadap susun

atur petua GK. Sebanyak tiga set susun atur petua GK telah dikenalpasti bagi penandaan GK ini seperti yang ditunjukkan dalam Jadual 6, Jadual 7 dan Jadual 8. Setelah susun atur kedudukan GK bagi ketiga-tiga set penandaan tersebut dimuktamadkan, pengujian dilakukan ke atas ketiga-tiga set tersebut dengan menggunakan ukuran kejituan yang dinilai dalam bentuk peratusan.

JADUAL 6. Susunan petua GK bagi set 1

Kedudukan Petua	Jenis GK	Kedudukan Petua	Jenis GK
1.	Kata Sendi	16.	Kata Tanya
2.	Kata Hubung	17.	Kata Penentu
3.	Kata Adverba	18.	Kata Perintah
4.	Kata Pemeri	19.	Kata Bilangan
5.	Penanda Wacana	20.	Simbol
6.	Kata Penegas	21.	Kata Nama Khas
7.	Ganti Nama Relatif	22.	Penjodoh Bilangan
8.	Kata Arah	23.	Gelaran
9.	Kata Ganti Nama	24.	Kependekan
10.	Kata Penguat	25.	Kata Pinjaman
11.	Kata Bantu Ragam	26.	Kata Kerja
12.	Kata Bantu Aspek	27.	Kata Adjektif
13.	Kata Nafi	28.	Kata Nama
14.	Kata Pembenar	29.	Kata Penekan
15.	Kewujudan	30.	Kata Pembenda

JADUAL 7. Susunan petua GK bagi set 2

Kedudukan Petua	Jenis GK	Kedudukan Petua	Jenis GK
1.	Kata Bilangan	16.	Kata Kerja
2.	Kata Hubung	17.	Kata Adjektif
3.	Kata Sendi	18.	Kata Penekan
4.	Penanda Wacana	19.	Kata Pembenda
5.	Kata Bantu Ragam	20.	Kata Penegas
6.	Kata Bantu Aspek	21.	Kata Arah
7.	Kata Adverba	22.	Kata Penentu
8.	Kata Penguat	23.	Kata Tanya
9.	Kata Ganti Nama	24.	Kata Perintah
10.	Ganti Nama Relatif	25.	Penjodoh Bilangan
11.	Kata Pemeri	26.	Kewujudan
12.	Kata Nafi	27.	Kata Pinjaman
13.	Kata Pembenar	28.	Gelaran
14.	Kata Nama Khas	29.	Kependekan
15.	Kata Nama	30.	Simbol

JADUAL 8. Susunan petua GK bagi set 3

Kedudukan Petua	Jenis GK	Kedudukan Petua	Jenis GK
1.	Kata Bilangan	16.	Kata Nama Khas
2.	Simbol	17.	Kata Nafi
3.	Kependekan	18.	Kata Pembenar
4.	Kata Penegas	19.	Kewujudan
5.	Gelaran	20.	Kata Tanya
6.	Kata Adverba	21.	Kata Penentu
7.	Kata Sendi	22.	Penjodoh Bilangan
8.	Ganti Nama Relatif	23.	Kata Pinjaman
9.	Kata Pemeri	24.	Kata Ganti Nama
10.	Kata Hubung	25.	Kata Arah
11.	Kata Penguat	26.	Kata Adjektif
12.	Kata Bantu Ragam	27.	Kata Kerja
13.	Kata Bantu Aspek	28.	Kata Nama
14.	Penanda Wacana	29.	Kata Penekan
15.	Kata Perintah	30.	Kata Pembenda

SENARIO PENGUJIAN

Senario pengujian perlu dilakukan bagi memastikan proses yang berlaku pada setiap uji kaji dilakukan secara tersusun. Tujuan uji kaji ini dijalankan ialah bagi mendapatkan peratusan nilai kejituan di antara ketiga-tiga set susun atur petua GK. Kemudian, salah satu set susun atur GK tersebut akan dibandingkan dengan penanda GK piawaian emas yang telah dibangunkan oleh Alfred et al. (2013). Secara umumnya, nilai peratusan kejituan bergantung susunan aturan petua GK dan jumlah petua GK yang digunakan.

Empat langkah yang dikenal pasti bagi menjalankan uji kaji yang dinyatakan adalah seperti berikut:

Langkah 1: Penyediaan korpus ujian di mana sebanyak 20 korpus daripada 100 korpus digunakan yang terdiri daripada 3,879 token termasuk tanda bacaan dan simbol.

Proses penyediaan korpus bermula dengan pengumpulan dan pemilihan korpus menggunakan data sekunder yang terdiri daripada teks mentah. Korpus yang dipilih dari sumber Berita Harian tersebut diperoleh secara atas talian meliputi pelbagai domain dan melalui proses pra-pemprosesanseperti yang telah dinyatakan sebelum ini. Hasil daripada pra-pemprosesan tersebut menjadi korpus yang tidak bertanda yang akan digunakan dalam pembangunan dan pengujian dalam kajian ini. Sebanyak 80 korpus dipilih sebagai korpus latihan yang mengandungi 16,548 token termasuk tanda bacaan dan simbol dan 20 korpus dijadikan korpus pengujian yang mengandungi 3,879 token termasuk tanda bacaan dan simbol yang menjadikan jumlah korpus kesemuanya sebanyak 100 korpus (20,427 token termasuk tanda bacaan dan simbol). Sebanyak 17 domain yang terlibat dalam kajian ini dan taburan domain-domain tersebut adalah seperti di Jadual 9:

JADUAL 9. Taburan domain-domain yang terlibat

Bil.	Jenis Domain	Jumlah Korpus	
		Korpus Latihan	Korpus Pengujian
1.	Agama Islam	5	2
2.	Alam Sekitar	5	1
3.	Bencana Alam	7	1
4.	Biologi	2	1
5.	Ekonomi	4	1
6.	Kemalangan	8	1
7.	Kemiskinan	5	1
8.	Kes Jenayah	12	3
9.	Kesenian	3	1
10.	Pelancongan	4	1
11.	Penternakan	2	1
12.	Penyakit	5	1
13.	Perdagangan	5	1
14.	Perhutanan	1	1
15.	Perniagaan	5	1
16.	Pertanian	3	1
17.	Teknologi Maklumat	4	1
	Jumlah	**80**	**20**

Kamus penandaan GK juga dibina secara manual dengan memasukkan senarai perkataan yang terdiri daripada kata akar sahaja yang dengan penanda GK masing-masing bagi membentuk leksikon. Penggunaan kamus ini adalah untuk memadankan perkataan kata akar yang mempunyai satu penanda GK sahaja.

Langkah 2: Mendapatkan nilai kejituan dalam bentuk peratusan bagi ketiga-tiga set susunan aturan petua GK.

Penilaian yang dilakukan adalah bagi perkataan yang mempunyai lebih daripada satu penanda GK iaitu menggunakan senarai petua dan hubungan kata. Keputusan pengujian dinilai dari segi kejituan penandaan dan direkodkan dalam jadual. Kejituan perbandingan dihasilkan dalam bentuk peratusan dengan menggunakan formula berikut:

$$Kejituan = \frac{bilangan\ perkataan\ yang\ ditanda\ dengan\ betul}{jumlah\ \ keseluruhan\ perkataan\ yang\ ditanda}\ x\ 100\%$$

Formula ini merupakan formula yang digunakan oleh Hassan (2015) dalam kajian beliau. Formula ini digunakan semula dalam kajian ini bagi mengukur ketepatan penandaan GK kerana mengambilkira bilangan perkataan yang hanya ditandakan dengan betul oleh sistem yang dibangunkan dan kemudiannya dibahagikan dengan jumlah keseluruhan perkataan termasuk simbol dan tanda bacaan yang terdapat dalam satu korpus yang ditandakan secara manual.

Langkah 3: Membandingkan hasil keputusan di antara ketiga-tiga set susunan aturan petua GK.

Keputusan nilai kejituan antara ketiga-tiga set susunan aturan petua GK dibandingkan dan set yang memperoleh nilai peratusan kejituan yang tertinggi dipilih untuk dibandingkan pula dengan keputusan penanda GK garis asas.

Langkah 4: Membina kesimpulan daripada hasil perbandingan tersebut.

Hasil keputusan yang didapati daripada perbandingan antara ketiga-tiga set susunan aturan petua GK dan perbandingan antara penanda GK yang dibina dalam kajian ini dengan penanda GK garis asas direkodkan dalam jadual untuk tujuan analisis dan seterusnya kesimpulan dibuat berdasarkan hasil analisis dan perbincangan.

PERBINCANGAN

Berdasarkan uji kaji yang dilakukan, perbincangan yang menyeluruh dilakukan bagi merumus sama ada penanda GK yang diguna dalam kajian adalah baik atau sebaliknya dalam menangani masalah seperti yang dikenal pasti. Contoh keputusan yang yang diperoleh daripada uji kaji dengan nilai kejituan adalah seperti berikut:

1. Analisis ketiga-tiga set susunan aturan petua GK

Analisis untuk nilai kejituan antara ketiga-tiga set aturan petua GK ini dilakukan dengan mengguna sejumlah 20 korpus daripada 100 korpus ayat yang dipilih dan kemudian dianalisis. Contoh kejituan yang diperoleh daripada analisis 20 korpus adalah seperti Jadual 10 berikut:

JADUAL 10. Perbandingan keputusan di antara ketiga-tiga set aturan petua GK

Korpus ujian	**Kejituan (%)**		
	Set 1	**Set 2**	**Set 3**
1	89.56	82.76	90.80
2	86.94	88.53	93.42
3	88.81	87.75	91.27
4	88.78	86.11	94.00
5	89.68	87.85	91.04
6	90.76	86.92	93.03
7	89.78	86.64	92.09
8	91.87	89.76	87.75
9	88.79	85.87	88.46
10	88.68	85.25	93.17
11	90.05	84.79	92.74
12	86.98	85.21	94.64
13	88.76	86.97	95.71
14	90.86	83.25	93.03
15	87.27	83.34	92.82
16	86.97	84.23	97.34
17	87.34	89.07	95.21
18	87.96	84.22	94.15
19	87.64	88.07	93.14
20	88.96	84.51	97.34
Purata (%)	**88.82**	**86.06**	**93.06**

Keputusan yang diperoleh, menunjukkan nilai kejituan bagi set 1, set 2 dan set 3, adalah sebanyak 88.82%, 86.06% dan 93.06%. Perbezaan hasil keputusan tersebut adalah disebabkan oleh perbezaan susun atur petua GK bagi setiap set. Set 1 dan set 3 mempunyai susunan aturan petua GK yang sama dari segi kedudukan GK KN, KPT dan KPB. Ini menunjukkan kedudukan GK KN, KPT dan KPB seharusnya berada di kedudukan terakhir kerana perkataan dalam korpus banyak merujuk kepada GK KN ini sama ada secara petua imbuhan mahupun secara

penandaan lalai. Idea meletakkan GK KN, KA, KK seterusnya KPT dan KPB di tengah-tengah kurang tepat kerana penandaan GK bagi set 2 paling rendah di antara ketiga-tiga set ini. Memandangkan aturan bagi KN disusun mendahului KA dan KA, maka setiap perkataan yang mempunyai imbuhan apitan *ke-…-an* kebanyakannya ditanda dengan GK ini. Begitu juga dengan perkataan yang bermula dengan imbuhan awalan *pe-* yang kebanyakannya turut ditandakan dengan GK KN. Justeru, berdasarkan keputusan kejituan secara keseluruhan, set 3 telah mencapai keputusan prestasi kejituan yang tertinggi dan dipilih sebagai penandaan petua GK yang terbaik bagi melakukan perbandingan dengan penanda GK garis asas.

2. Analisis set 3 dan penanda GK garis asas

Perbincangan yang menyeluruh dilakukan bagi merumuskan sama ada penanda GK yang diguna dalam kajian adalah baik atau sebaliknya dalam menangani masalah seperti yang dikenal pasti dalam kajian. Analisis untuk nilai kejituan antara set 3 dengan penanda GK garis asas dilakukan dengan menggunakan 20 korpus yang sama. Hanya petua imbuhan bagi KN, KK dan KA serta petua hubungan kata daripada kajian Alfred et al. (2013) yang digunakan sebagai piawaian emas dalam kajian ini. Contoh kejituan yang diperoleh daripada analisis 20 korpus adalah seperti Jadual 11 berikut:

JADUAL 11. Perbandingan keputusan keseluruhan antara dua penanda GK

Korpus ujian	**Jumlah token**	**Token yang dikira**	**Kejituan (%)**	
			Penanda GK kajian ini	**Penanda GK Alfred et al. (2013)**
1	174	104	90.80	78.26
2	228	129	93.42	77.64
3	229	142	91.27	76.08
4	200	120	94.00	77.87
5	201	127	91.04	78.11
6	201	117	93.03	79.14
7	177	106	92.09	78.62
8	204	132	87.75	73.53
9	208	131	88.46	76.04
10	205	126	93.17	75.96
11	179	108	92.74	73.93
12	168	111	94.64	70.93
13	210	141	95.71	79.90
14	201	113	93.03	78.94
15	181	111	92.82	79.17
16	188	113	97.34	78.74
17	182	117	95.21	73.94
18	180	116	94.15	78.87
19	175	106	93.14	78.29
20	188	130	97.34	79.40
Purata Keseluruhan (%)			**93.06**	**77.17**

Jumlah token adalah jumlah kesemua perkataan termasuk simbol dan tanda bacaan yang terdapat dalam satu korpus manakala jumlah token yang dikira adalah jumlah perkataan dan simbol unik dengan tidak mengambilkira perkataan sama yang diulang atau digunakan dalam korpus tersebut. Berdasarkan keputusan uji kaji yang ditunjukkan dalam Jadual 11, set 3 mempunyai nilai kejituan sebanyak 93.06% berbanding dengan nilai kejituan garis asas sebanyak 77.17%, yang menunjukkan perbezaan hasil kejituan yang ketara iaitu 15.89%. Bagi

penandaan GK garis asas, terdapat beberapa ralat pada penandaan perkataan. Misalnya perkataan yang sepatutnya ditandakan sebagai KNK tetapi ditandakan sebagai KK seperti *Maluri, Malaysia* dan *Terengganu* manakala perkataan *Seremban* pula ditandakan sebagai KA. Selain itu, perkataan *mereka* yang sepatutnya ditandakan sebagai KGN, ditandakan sebagai KK. Ralat-ralat ini terjadi disebabkan faktor kekurangan petua bagi sesetengah GK seperti KNK, KGN, TEN, KPM, GNR, ADA dan petua-petua GK yang lain seperti yang telah ditunjukkan dalam Jadual 2.

Manakala bagi penandaan GK kajian ini juga, terdapat beberapa ralat pada penandaan perkataan seperti bibir, buah, hikmah, ibadah, ijazah, jenayah yang sepatutnya ditandakan sebagai KN tetapi ditandakan sebagai KA. Ini disebabkan oleh susunan aturan petua KA yang disusun terlebih dahulu daripada KN selain petua KA yang menetapkan syarat bahawa pemadanan bagi GK ini adalah bagi perkataan yang bermula dengan imbuhan awalan *ter*, *se*, *bi* atau perkataan yang tidak bermula dengan *di* dan *men* tetapi berakhir dengan *–ah.* Begitu juga dengan perkataan *enggan*, *mapan* dan *sistematik* yang sepatutnya ditandakan sebagai KA tetapi ditandakan sebagai KN kerana syarat bagi penandaan GK KN adalah perkataan yang mempunyai imbuhan akhiran *–an* atau *–ik.*

KESIMPULAN

Kajian ini membangunkan petua baru bagi penandaan GK bahasa Melayu dan membandingkan penanda GK yang dibangunkan dalam kajian ini dengan prestasi penandaan GK bahasa Melayu berasaskan petua dengan piawaian emas sedia ada. Sebanyak 15 petua GK baru telah berjaya ditambah daripada 15 petua sedia ada yang menjadikan keseluruhan petua penandaan GK sebanyak 30 petua. Bagi hubungan kata pula, sebanyak dua petua baru ditambah daripada 14 petua sedia ada yang menjadikan jumlah petua bagi hubungan kata sebanyak 16 petua. Penyusunan dan aturan kedudukan petua dalam penandaan GK juga amat penting kerana susun atur kedudukan jenis GK yang kurang tepat akan menjejaskan prestasi penandaan GK terhadap sesuatu perkataan. Keputusan susun atur petua GK mendapati kedudukan GK KN, KPT dan KPB seharusnya berada di kedudukan terakhir kerana perkataan dalam korpus banyak merujuk kepada GK KN ini sama ada secara petua imbuhan mahupun secara penandaan lalai. Hasil uji kaji menunjukkan keputusan kejituan penanda GK yang dibangunkan dalam kajian ini adalah lebih baik dengan nilai kejituan 93.06% berbanding nilai kejituan piawaian emas iaitu sebanyak 77.17%. Ini menunjukkan bahawa susunan aturan kedudukan petua GK dan jumlah petua GK yang digunakan mempengaruhi nilai kejituan penandaan GK. Hasil kajian ini dapat membantu para penyelidik dalam melaksanakan penandaan golongan kata bagi korpus bahasa Melayu dengan menghasilkan nilai ketepatan yang lebih tinggi melalui penambahan petua baru. Sebagai penambahbaikan dalam kajian masa hadapan, penggunaan perkataan majmuk perlu diambilkira kerana kebanyakan perkataan ini digunakan dalam kebanyakan sumber berita. Selain itu juga, korpus daripada sumber sumber media sosial boleh digunakan kerana walaupun bahasa yang digunakan bagi sumber ini kebanyakannya tidak formal dan berdepan dengan isu data hingar namun kandungan maklumat yang disebarkan melalui media sosial ini adalah cepat dan terkini.

RUJUKAN

Abdullah, H. 2006. Morfologi : Siri Pengajaran dan Pembelajaran Bahasa Melayu. PTS Publications and Distributors Sdn.Bhd.

Alfred, R., Mujat, A. & Obit, J. H. 2013. A Ruled-Based Part of Speech (RPOS) tagger for Malay text articles 7803 LNAI(PART 2), 50–59.

Haroon, A. R. 2010. Bab Empat : Metodologi kajian. Universiti Malaya.

Hassan, M. 2015. Model Markov Tersembunyi Tak Terselia untuk Penandaan Golongan Kata Bahasa Melayu. Universiti Kebangsaan Malaysia.
Juhaida, A. B., Khairuddin, O., Mohammad Faidzul, N. & Mohd Zamri, M. 2013. Morphology analysis in Malay POS Prediction. Proceeding of the International Conference on Artificial Intelligence in Computer Science and ICT 2013, (November), 25–26.
Jurafsky, D. & Martin, J. H. 2011. Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition hlm.Second Edi. Pearson Prentice Hall.
Kumawat, D. & Jain, V. 2015. POS Tagging Approaches : A Comparison 118(6), 32–38.
Mohamed, L. & Rohana, M. 2015. Building A Dictionary of Malay Language Part-of-Speech Tagged Words Using Bahasa WordNet and Bahasa Indonesia Resources 1–8.
Mohd Pouzi, H. & Syarifah Fatem Na'imah, S. K. 2014. Part of Speech Tagger for Malay Language Based on Word Morphology 2014(October), 1499–1502.
Nik Safiah, K., Farid M., O., Hashim, H. M. & Abdul Hamid, M. 2015. Tatabahasa Dewan Edisi Ketiga 21, 23–25,43.
Preeti & Sidhu, B. K. 2013. Natural Language Processing. Encyclopedia of Systems Biology, 4(5), 751–758.
Rashel, F., Luthfi, A., Dinakaramani, A. & Manurung, R. 2014. Building an Indonesian rule-based part-of-speech tagger. Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014, 70–73.
Widhiyantil, K. & Agus, H. 2012. POS Tagging Bahasa Indonesia Dengan HMM dan Rule Based. Informatika, 8(2).
Zuraidah, M. D. 2010. Processing Natural Malay Texts: A Data-driven Approach. Trames, 14(1), 90–103.

Nur Ashikin Halid
Nazlia Omar
Fakulti Teknologi dan Sains Maklumat (FTSM),
UKM, Bangi, 43600, Selangor, Malaysia.
ashikin.halid@gmail.com, nazlia@ukm.edu.my

Received: 16 June 2017
Accepted: 19 December 2017