PENANDAAN GOLONGAN KATA BAHASA MELAYU MENGGUNAKAN PENDEKATAN BERASASKAN PETUA

NUR ASHIKIN BINTI HALID

UNIVERSITI KEBANGSAAN MALAYSIA

PENANDAAN GOLONGAN KATA BAHASA MELAYU MENGGUNAKAN PENDEKATAN BERASASKAN PETUA

NUR ASHIKIN BINTI HALID

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA TEKNOLOGI MAKLUMAT

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2017

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang setiap satunya telah saya jelaskan sumbernya.

21 Jun 2017

NUR ASHIKIN BINTI HALID
GP04131

PENGHARGAAN

Syukur Alhamdulillah kepada Allah S.W.T kerana dengan limpah kurniaan-Nya telah memberikan saya kesihatan yang baik, masa dan kematangan fikiran bagi menyiapkan kajian ini. Jutaan terima kasih yang tidak terhingga kepada penyelia saya, Prof. Madya Dr. Nazlia Omar yang telah banyak memberi bimbingan, tunjuk ajar, teguran dan nasihat yang begitu berguna sepanjang kajian ini. Tidak dilupakan juga kepada penyelaras program Prof. Madya Dr. Kamsuriah Ahmad yang turut membantu memberikan pandangan dalam menyempurnakan kajian ini.

Ucapan terima kasih yang tidak terhingga juga ditujukan kepada ahli keluarga tercinta khususnya suami saya Mohd Imran Md. Yusop, ibu saya Khalidjah Darus dan Rohanah Selamat, bapa saya Halid Ahmad, adik saya Asyraf Hadafi bin Halid serta adik-beradik saya yang lain atas segala doa, pengorbanan, dorongan dan kesabaran yang diberikan sepanjang saya menyiapkan kajian ini. Ucapan penghargaan ini juga ditujukan kepada rakan-rakan seperjuangan yang banyak memberikan tunjuk ajar, nasihat dan motivasi dalam mengharungi cabaran-cabaran sepanjang pengajian ini.

Penghargaan dan terima kasih kepada Jabatan Perkhidmatan Awam yang telah memberikan cuti penuh bagi saya melanjutkan pengajian ini. Akhir bicara, saya mengucapkan terima kasih kepada mereka yang terlibat secara langsung atau tidak langsung sehingga terhasilnya tesis ini.

ABSTRAK

Penandaan Golongan Kata (GK) merupakan satu 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. Sebagai sebahagian daripada tugas asas dalam Capaian Maklumat, proses ini adalah satu tugasan penting bagi prestasi pemprosesan teks. Masalah yang sering timbul dalam proses penandaan GK adalah kewujudan perkataan kabur dan perkataan yang tidak diketahui. Manakala masalah utama dalam penandaan GK bahasa Melayu adalah kekurangan petua dalam kajian sedia ada. Objektif utama kajian ini adalah untuk 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, pra-pemprosesan, pembangunan petua dan penilaian. Pengumpulan dan pemilihan korpus menggunakan data sekunder yang diperoleh daripada Berita Harian secara atas talian meliputi pelbagai domain. Sebanyak 100 korpus telah dipilih dan 80 daripadanya dijadikan sebagai korpus latihan manakala baki selebihnya sebagai korpus ujian. 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 keberkesanan 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%. Hasil daripada kajian ini diharap dapat membantu para penyelidik dalam melaksanakan penandaan golongan kata bagi korpus bahasa Melayu dengan menghasilkan nilai ketepatan yang lebih tinggi melalui penambahan petua baru.

MALAY PART OF SPEECH TAGGING USING RULED-BASED APPROACH

ABSTRACT

Part of Speech tagging (POS) is a process of annotating or assigning a tag in a sentence for each token or word as noun, verb, adjectives or adverb depending on the relationship of words and sentence definitions. As part of the basic tasks in Information Retrieval, this process has important function in text processing performance. Among the problems 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. The main objective of this research is 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, pre-processing, development and evaluation. The process of collection and selection of the corpus is using secondary data, obtained from online daily news which covers several domains. The total number of corpus used is 100, in which 80 of them have been used as a training corpus whiles the rest as a test corpus. 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%. Outcome from this research is hope to help future researchers in tagging Malay corpus by generating higher precision through the addition of the new rules.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

DBP	Dewan Bahasa dan Pustaka
EM	Entropi Maksimum
GK	Golongan Kata
MEMM	Model Entropi Maksimum Markov
MMT	Model Markov Tersembunyi
MPK	Model Pepohon Keputusan
MVS	Mesin Vektor Sokongan
NMP	Nyahkabur makna-perkataan
PBI	Pembelajaran Berasaskan Ingatan
PBT	Pemprosesan Bahasa Tabii
PM	Pembelajaran Mesin
RPOS	*Rule-based Part of Speech*

BAB I

PENDAHULUAN

1.1 PENGENALAN

Pemprosesan Bahasa Tabii (PBT) atau *Natural Language Processing* (NLP) merujuk kepada sistem komputer yang cuba untuk memahami, menganalisa dan menghasilkan satu atau lebih bahasa yang digunakan oleh manusia *(human language)* (James 2003). Ia dijalankan secara berperingkat dengan dengan fasa pertama bermula pada lewat tahun 1940-an yang memberi tumpuan kepada Penterjemahan Mesin (*Machine Translation*). Penyelidikan dalam bidang ini seterusnya berkembang pada fasa kedua iaitu pada lewat tahun 1960-an, fasa ketiga pada lewat tahun 1970-an hingga 1980-an, fasa keempat pada tahun 1990-an sehingga kini (Jones 2001).

Terdapat beberapa cabang tugas utama atau kajian yang lazimnya dilaksanakan dalam bidang ini selain Penterjemahan Mesin, seperti Capaian Maklumat (*Information Retrieval*), Soal-Jawab (*Question - Answer*) (Liddy 2001), Pengecaman Entiti Nama (*Named Entity Recognition*), Pelabelan Peranan Semantik (*Semantic Role Labeling*) dan Nyahkabur Makna-Perkataan (*Word-Sense Disambiguation*) (Collobert et al. 2011). Sebagai sebahagian daripada tugas asas dalam Capaian Maklumat (IR), penandaan Golongan Kata (GK) atau Penandaan Tatabahasa merupakan satu tugasan penting bagi prestasi pemprosesan teks (Karimpour et al. 2008). Ia adalah satu 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, Mujat & Obit 2013a).

Sebelum ini, terdapat banyak kajian penyelidikan telah dilakukan dalam bidang penandaan GK yang melibatkan pelbagai bahasa, antaranya Bahasa Inggeris

(Brants 2000; Brill 1992; Marcus, Santorini & Marcinkiewicz 1993), Bahasa Cina (Chang & Chen 1993; Lua 1996; Ng & Low 2004), Bahasa Arab (Khoja 2001; Maamouri & Bies 2002), Bahasa Siam (Hansakunbuntheung, Tesprasit & Sornlertlamvanich 2003; Sornlertlamvanich, Charoenporn & Isahara 1997), Bahasa India (Kumar & Josan 2010), Bahasa Bengali (Ekbal, Haque & Bandyopadhyay 2007) dan Bahasa Indonesia (Pisceldo, Adriani & Manurung 2009; Wicaksono & Purwarianti 2010).

Bagi bahasa Eropah seperti Bahasa Inggeris, penandaan GK telah digunakan secara meluas dan digunakan melalui pelbagai pendekatan. Ini mungkin kerana kaedah-kaedah frasa tatabahasa yang mudah dan tidak rumit untuk memahami dan untuk digunakan (Alfred, Mujat & Obit 2013). Tambahan pula, majoriti alat PBT adalah berdasarkan kepada Bahasa Inggeris dan umumnya digunakan dalam penyelidikan (Alfred et al. 2013). Ia adalah lebih mencabar untuk bahasa Asia khususnya bahasa Melayu kerana beberapa perkataan Melayu mempunyai unsur pengubahsuaian daripada bahasa lain seperti bahasa Inggeris dan mempunyai perkataan tambahan (imbuhan, sama ada awalan, akhiran atau apitan) daripada perkataan asal mereka (Alfred et al. 2013).

1.2 LATAR BELAKANG KAJIAN

Terdapat beberapa teori mengenai asal usul Bahasa Melayu oleh beberapa pengkaji. Menurut Othman (1996), perkataan Melayu berasal daripada bahasa Sanskrit yang ditulis sebagai *Mo-lo-yeu*, dan orang *Mo-lo-yeu* tersebut dipercayai berasal dari Jambi, sebuah kerajaan yang terletak di Indonesia ketika itu (Ku Hasnita, Adlina & Mohd Hafiz 2013). Bahasa Melayu ini juga pernah menjadi *lingua franca*, iaitu bahasa perantaraan atau bahasa penghubung yang dituturkan oleh orang-orang yang berbeza bahasa asal bahasa peribuminya pada zaman kegemilangan empayar kerajaan Melayu Melaka pada zaman dahulu (Sew 2013).

Hasil daripada pertembungan bahasa pada zaman dahulu, bahasa Melayu asal mengalami beberapa perubahan dan kini, bahasa Melayu mempunyai unsur-unsur campuran dan pinjaman kata daripada bahasa lain seperti Arab, Cina dan India. Sebagai bahasa yang turut diangkat sebagai bahasa kebangsaan dan bahasa rasmi

negara terutama di Malaysia, Indonesia dan Brunei, Bahasa Melayu juga melalui peringkat pengembangan dan kemajuannya di mana usaha-usaha penyelarasan dan pemodenan giat dijalankan oleh Dewan Bahasa dan Pustaka yang berfungsi sebagai badan perancang bahasa (Awang 2010). Pengekodan Bahasa Melayu merupakan salah satu usaha yang dijalankan bagi menyeragamkan bahasa ini dalam bentuk yang sempurna dari aspek tatabahasa, ejaan, makna dan sebutan (Nik Safiah et al. 2015).

Terdapat dua bahagian dalam tatabahasa iaitu sintaksis yang merupakan bahagian pembentukan frasa dan ayat, dan morfologi; iaitu satu bidang ilmu yang mengkaji perkataan dari segi struktur, bentuk dan penggolongan kata. Sebagai bahasa yang kaya dengan kosa katanya, bahasa Melayu mempunyai morfologinya yang tersendiri bagi menghasilkan perkataan lain yang mempunyai makna selain kata akar yang memberi kesan kepada golongan katanya (H. Mohamed, Omar & Ab Aziz 2011)

Secara umumnya, penandaan GK melibatkan dua kaedah pembelajaran iaitu pembelajaran terselia dan tidak terselia. Pembelajaran terselia merupakan pembelajaran yang telah dikaji dengan beberapa kaedah yang telah mencapai ketepatan hampir dengan penandaan GK yang dilakukan oleh manusia (Li, Graça & Taskar 2012). Bagi kaedah ini, model perlu ditandakan terlebih dahulu kerana ianya akan digunakan untuk mempelajari maklumat mengenai set penandaan dan set petua bagi korpus latihan (Kumawat & Jain 2015). Manakala bagi kaedah pembelajaran tidak terselia pula, tiada sebarang penetapan kategori penanda GK mahupun teks yang telah ditanda dengan penanda GK yang diperlukan (Biemann, Giuliano & Gliozzo 2007). Sungguhpun begitu, kaedah pembelajaran terselia memerlukan data bertanda pada skala yang besar dan melibatkan penggunaan masa dan tenaga yang banyak dalam proses penyediaan keputusan data set berbanding dengan kaedah pembelajaran tidak terselia yang tidak memerlukan penyediaan data set atau model keputusan yang sebenar semasa latihan dijalankan (Faralli & Navigli 2012).

Terdapat tiga kategori utama bagi pendekatan yang telah digunakan dalam penandaan GK iaitu berasaskan linguistik, kaedah statistikal dan pendekatan pembelajaran mesin (Hasan, Uzzaman & Khan 2007; Kumawat & Jain 2015; Marquez, Padro & Rodriguez 1998). Contoh penandaan GK berasaskan linguistik

adalah pendekatan berasaskan petua dan penanda GK paling popular yang masih digunakan sebagai garis panduan sehingga kini adalah penanda nama Brill (Brill 1992). Contoh kaedah statistikal adalah Model Markov Tersembunyi (MMT), Entropi Maksimum (EM) dan pembelajaran berasaskan Transformasi (Hassan, Nazlia & Mohd Juzaidin 2014) manakala contoh kaedah pendekatan pembelajaran mesin yang merupakan kaedah gabungan pendekatan berasaskan petua dan kaedah statistikal adalah Pembelajaran Berasaskan Ingatan (PBI) dan Model Pepohon Keputusan (MPK).

Pendekatan berasaskan petua merupakan salah satu pendekatan terawal yang menggunakan pengetahuan linguistik untuk memberikan tag yang betul kepada setiap perkataan dalam ayat atau fail dengan menggunakan satu set petua yang ditulis (Kumawat & Jain 2015; Mubarak, Madhu & Shanavas 2015). Selain daripada mudah digunakan kerana ia menggunakan petua yang ringkas dan menjimatkan ruang, (Mubarak, Madhu & Shanavas 2015) penambahbaikan ke atas penandaan GK juga mudah dilakukan kerana pendekatan ini menggunakan set kecil petua yang mudah dan kurang kompleks (Brill 1992).

Pendekatan menggunakan kaedah statistikal pula menggunakan korpus latihan untuk memilih tag yang paling mungkin untuk suatu perkataan (Hasan, Uzzaman & Khan 2007). Ia memberi keutamaan kepada kaedah kebarangkalian dan maklumat statistik berbanding peraturan tatabahasa. Teknik MMT merupakan teknik yang paling luas digunakan dalam pendekatan ini kerana ianya lebih ringkas dan lebih tepat berbanding pendekatan lain jika korpus bertanda wujud dalam jumlah atau skala yang besar (Hassan, Nazlia & Mohd Juzaidin 2014), selain dapat menanda perkataan dalam jumlah yang banyak dalam masa yang singkat (Al-Shamsi et al. 2006).

Sementara itu, pendekatan pembelajaran mesin pula merupakan pendekatan yang mengambil tag yang paling mungkin berdasarkan korpus latihan dan mengaplikasikan satu set petua tertentu untuk melihat sama ada suatu tag tersebut perlu ditukar kepada tag yang baru atau dikekalkan (Hasan, Uzzaman & Khan 2007). Misalnya menggunakan teknik PBI, ianya adalah berasaskan kepada andaian bahawa penaakulan adalah berdasarkan kepada penggunaan semula secara langsung dari

pengalaman yang disimpan, bukan pada penggunaan pengetahuan (Hasan et al. 2007). Sehubungan itu, pembelajaran menjadi lebih cepat dan berkeupayaan memberikan penjelasan dengan lebih baik terhadap hasil penandaan GK tersebut (Daelemans et al. 1996).

1.3 PENYATAAN MASALAH

Menurut Dhanalakshmi et al. (2009), masalah utama yang sering berlaku dalam proses penandaan GK adalah kewujudan perkataan kabur (*ambiguous*) dan perkataan yang tidak diketahui (*unknown*). Perkataan kabur adalah perkataan yang mempunyai lebih daripada satu makna atau boleh mempunyai lebih daripada satu tag. Dalam bidang linguistik, istilah ini dikenali sebagai homonim dan homograf. Menurut Nik Safiah (2003), homonim merupakan perkataan yang mempunyai bunyi yang serupa tetapi membawa maksud yang berlainan. Misalnya perkataan *pukul* yang boleh mempunyai dua makna yang berbeza. Ia boleh dikategorikan sebagai Kata Nama dan Kata Kerja seperti yang digunakan dalam ayat-ayat berikut : "Rombongan sekolah itu dijangkakan akan tiba pada pukul dua petang" dan ""Ali, jangan pukul kucing itu, kata ibu kepada Ali"". Manakala homograf pula adalah perkataan yang mempunyai ejaan yang sama tetapi berbeza sebutan dan maknanya. Misalnya perkataan *perang* yang membawa maksud sejenis warna dan pertempuran. Bagi manusia, tugas meletakkan tag bagi kedua-dua ayat tersebut adalah mudah kerana manusia mempunyai akal pengetahuan dan naluri yang boleh memproses maklumat dan dapat mengenalpasti perbezaan makna perkataan yang digunakan dalam kedua-dua ayat tersebut (Kumawat & Jain 2015).

Sementara itu, perkataan yang tidak diketahui atau perkataan baru atau perkataan anu pula adalah perkataan yang sama ada tidak terdapat dalam korpus latihan ataupun tidak terdapat dalam kamus penandaan GK. Lazimnya, jenis perkataan yang tidak diketahui ini adalah disebabkan oleh salah ejaan, kata nama khas, kata singkatan, kata pinjaman atau perkataan dari bahasa asing (Ranaivo-Malancon, Chua & Ng 2007). Seringkali perkataan ini menjadi kata kunci dalam sesuatu ayat bagi menterjemahkan maksud keseluruhan ayat tersebut. Sebagai manusia, kita boleh meneka maksud ayat tersebut berdasarkan pengetahuan dan pengalaman yang

diperoleh sebelum ini namun tidak bagi mesin. Suatu set petua perlu ditentukan apabila terdapat perkataan yang tidak diketahui dan tugasan menetapkan petua ini bukan suatu tugasan yang mudah.

Masalah utama dalam penandaan GK bahasa Melayu adalah 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. Ini kerana ia dianggap sebagai imbuhan yang tidak popular dan tidak produktif yang boleh meningkatkan masalah kekaburan dalam proses penandaan GK, berbanding dengan imbuhan lain seperti imbuhan awalan, akhiran dan apitan (Abdullah 2006). Tambahan pula dari sudut bahasa, sesuatu perkataan asal atau perkataan akar yang telah ditambah imbuhan, akan mengubah makna asal perkataan tersebut (Nik Safiah et al. 2015). Penambahan imbuhan juga menyebabkan perubahan kepada golongan kata sesuatu perkataan.

1.4 MATLAMAT DAN OBJEKTIF KAJIAN

Objektif kajian ini adalah seperti berikut :

i. Membangunkan petua baru bagi penandaan GK Bahasa Melayu.
ii. Membandingkan prestasi penandaan GK Bahasa Melayu berasaskan petua dengan piawaian emas sedia.

1.5 SKOP KAJIAN

Skop kajian ini tertumpu kepada sumber berita dalam Bahasa Melayu yang dipetik daripada Berita Harian secara atas talian mencakupi beberapa bidang seperti agama Islam, ekonomi, perdagangan, kemiskinan, kesenian, pertanian, kes jenayah, kemalangan, teknologi maklumat, biologi, alam sekitar, perniagaan, perhutanan, penternakan, pelancongan, penyakit dan bencana alam.

1.6 METOD KAJIAN

Metod kajian yang digunakan dalam kajian ini merangkumi rangka kerja penyelidikan yang memaparkan ringkasan keseluruhan proses kajian yang dijalankan bermula daripada proses pengumpulan dan pemilihan korpus, pra-pemprosesan, pembangunan petua dan penilaian.

Proses pengumpulan dan pemilihan korpus yang dijalankan dari awal permulaan kajian dengan menggunakan data sekunder yang dipetik daripada sumber Berita Harian yang didapati secara atas talian yang meliputi skop seperti yang diterangkan pada bahagian 1.5. Sebanyak 100 korpus telah dipilih dan 80 daripadanya dijadikan sebagai korpus latihan semasa pembangunan petua manakala selebihnya dijadikan sebagai korpus ujian semasa proses penilaian. Petua bagi setiap GK turut dibangunkan dan disusun mengikut aturan masing-masing. Aturan ini amat penting kerana ia memberi kesan terhadap hasil penandaan GK pada sesi pengujian kelak.

Proses pra-pemprosesan menerangkan bagaimana artikel berita dalam bentuk teks mentah melalui proses pemisahan ayat dan pentokenan bagi mendapatkan korpus yang tidak bertanda. Korpus ini akan dijadikan sebagai input dalam pembangunan petua GK. Dalam proses ini juga, kamus GK telah dibina bagi membentuk leksikon yang terdiri daripada kata akar bagi proses perkataan yamg memiliki satu jenis GK sahaja.

Proses pembangunan petua pula merupakan proses bagi mencirikan setiap jenis GK kepada petuanya yang tersendiri di samping menyusun dan mengatur kedudukan jenis GK dalam pembangunan petua itu sendiri. Sebanyak 30 petua GK dan 16 hubungan kata dibangunkan dalam proses ini.

Proses terakhir merupakan proses penilaian bagi melihat keberkesanan pembangunan petua yang telah dibangunkan dan penyusunan aturan petua yang terbaik bagi menghasilkan keputusan penandaan GK yang terbaik. Hasil pengujian pembangunan petua yang diukur melalui kejituan penandaan GK ini dibandingkan dengan hasil yang didapati daripada penanda GK piawaian emas.

1.7 ORGANISASI KAJIAN

Organisasi kajian bagi tesis ini terdiri daripada lima bahagian iaitu pendahuluan, kajian literasi, metod kajian, implementasi dan pengujian serta perbincangan dan kesimpulan.

Bab I dimulakan dengan pengenalan dan latar belakang kajian yang dijalankan. Seterusnya, pernyataan masalah bagi kajian ini dan objektif kajian yang ingin dicapai juga dinyatakan. Bab ini juga mengandungi skop kajian dan memaparkan metod kajian secara ringkas bagi kajian yang akan dilaksanakan.

Bab II membincangkan kajian literasi yang telah dijalankan dengan menggunakan pendekatan berasaskan petua bagi bahasa Melayu yang melibatkan teknik berlainan dalam kajian lepas. Perbandingan hasil kajian lepas dengan menggunakan pendekatan yang sama atau pendekatan berbeza turut dibincangkan dalam bab ini.

Bab III pula memperincikan metod kajian yang digunakan dalam kajian ini yang terdiri daripada pengumpulan dan pemilihan korpus, pra-pemprosesan, pembangunan petua dan penilaian. Proses menjadikan korpus daripada teks mentah kepada korpus tidak bertanda untuk dijadikan sebagai input dalam kajian turut dijelaskan. Pembangunan petua bagi 30 jenis GK dan 16 hubungan kata juga diperincikan dalam bab ini.

Bab IV merupakan fasa implimentasi dan perbincangan di mana pengujian dilakukan dengan membandingkan antara set susunan aturan petua GK yang dibangunkan dan juga perbandingan hasil keputusan daripada set susunan aturan petua GK yang dipilih dengan keputusan penandaan GK piawaian emas. Analisis keputusan secara terperinci turut dilakukan dalam bab ini.

Bab V merupakan bab terakhir dimana kesimpulan dibuat daripada kajian yang telah dijalankan serta memaparkan penemuan kajian dan kekangan yang terdapat dalam kajian ini untuk dijadikan sebagai penambahbaikan pada kajian yang akan datang serta cadangan penambahbaikan.

1.8 KESIMPULAN

Bab ini menerangkan secara menyeluruh tentang latar belakang kajian, pernyataan masalah tentang senario semasa yang sedang dihadapi, matlamat dan objektif kajian, skop kajian serta penerangan ringkas tentang metod kajian yang digunakan. Objektif bagi kajian ini adalah untuk membangunkan petua baru bagi penandaan GK bahasa Melayu berdasarkan piawaian emas sedia ada.

BAB II

KAJIAN LITERASI

2.1 PENGENALAN

Bab ini memaparkan kajian literasi terhadap kajian dan penyelidikan yang telah dilaksanakan oleh para penyelidik terdahulu dalam bidang penandaan GK khususnya dalam bahasa Melayu. Sebagaimana yang diterangkan sebelum ini, terdapat banyak kajian dalam bidang penandaan GK yang telah dilaksanakan dengan melibatkan pelbagai bahasa dan juga pendekatan yang berbeza. Prestasi pendekatan yang digunakan dalam satu proses penandaan GK dalam satu bahasa mungkin berbeza dengan satu bahasa yang lain kerana setiap bahasa mempunyai ciri-ciri dan tatabahasanya yang tersendiri. Secara amnya, bab ini terdiri daripada enam bahagian yang dimulai dengan penandaan GK, seterusnya kaedah penandaan GK, perkataan kabur dan perkataan tidak diketahui, GK bahasa Melayu, kajian lepas berkenaan penandaan GK dan diakhiri dengan kesimpulan bagi bab ini.

2.2 PENANDAAN GOLONGAN KATA

Seperti yang dibincangkan dalam Bab I, proses penandaan GK merupakan satu proses menganotasi tanda nama dalam ayat untuk setiap kelas token atau perkataan seperti kata nama, kata kerja, kata sifat dan kata keterangan bergantung kepada hubungan perkataan dan juga definisi ayat (Alfred et al. 2013). Kebanyakan penanda GK telah dilatih dari *treebanks* dalam domain agensi berita, seperti korpus *Wall Street Journal* dari korpus *Penn Treebank* (Gimpel et al. 2011).

Namun begitu, dengan mengambilkira keperluan dan ciri tatabahasa dalam bahasa Melayu, penanda GK yang digunakan dalam korpus bahasa Inggeris tidak sesuai untuk digunakan dalam korpus bahasa Melayu. Antara perbezaan yang ketara adalah tiada penggunaan penjodoh bilangan yang digunakan dalam bahasa Inggeris sebagaimana yang digunakan dalam bahasa Melayu (Hassan, Nazlia & Mohd Juzaidin 2014), tiada keperluan bagi penggunaan kata nama tunggal (*single*) atau jamak (*plural*) seperti yang digunakan dalam bahasa Inggeris dan perbezaan penggunaan imbuhan dalam perkataan akar serta perbezaan struktur morfologinya (Norsimah et al. 2007).

Dalam bidang kajian linguistik, terdapat tujuh kriteria yang digunakan bagi mengkategorikan perkataan ke dalam sesuatu golongan, iaitu kriteria fonologi, morfologi, sintaksis, leksikal, semantik, pragmatik dan wacana (*discourse*) (Norliza 2008; Preeti & Sidhu 2013). Namun begitu, kajian ini hanya memfokuskan kepada morfologi dan sintaksis sahaja 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. Akar merupakan bahagian utama morfem yang mengandungi makna penting (Juhaida et al. 2013) manakala imbuhan pula menghasilkan kata terbitan selain pemajmukan kata dan penggandaan kata yang bertujuan untuk menambahkan lagi perbendaharaan kosa kata bahasa Melayu (Imran Ho & Hazimah 2015).

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).

2.3 KAEDAH PENANDAAN GOLONGAN KATA

Penandaan Golongan Kata (GK) boleh dilaksanakan melalui tiga pendekatan iaitu pendekatan berasaskan petua (*rule-based approach*), pendekatan berasaskan statistikal (*statistical approach*) dan pendekatan pembelajaran mesin (PM) iaitu gabungan pendekatan petua dan statistikal (Kumawat & Jain 2015). Setiap pendekatan ini mempunyai kelebihan dan kekurangan yang tersendiri berdasarkan saiz korpus dan domain yang dipilih.

2.3.1 Pendekatan Berasaskan Petua

Pendekatan berasaskan petua atau peraturan menggunakan satu set peraturan yang bertulis untuk digunakan sebagai penanda GK bagi perkataan berdasarkan peraturan yang telah disediakan (Kumawat & Jain 2015). Peraturan –peraturan ini secara amnya dikenali sebagai peraturan kerangka konteks. Dua peringkat arkitektur digunakan dalam algoritma terawal untuk meletakkan penandaan GK secara automatik (Dalal et al. 2007). Pada peringkat awal, kamus digunakan bagi menandakan setiap perkataan dengan penanda yang paling hampir atau paling sesuai. Peringkat seterusnya adalah menggunakan petua-petua yang telah disetkan bagi menyahkabur perkataan (Kumawat & Jain 2015).

Penandaan GK yang menggunakan pendekatan ini menggunakan kamus penandaan GK dan peraturan mengenai imbuhan bagi mengenalpasti definisi sebenar perkataan tersebut. Dalam kajian ini, kaedah penandaan GK berasaskan petua oleh Brill (1992) dan Alfred et al. (2013) telah digabungkan dan digunakan. Sebahagian besar ciri-ciri yang digunakan adalah berdasarkan modul Brill (1992) yang bermula dengan memberikan penanda GK lalai (*default*) iaitu tag yang paling kerap muncul dalam sesuatu korpus kepada suatu perkataan. Dalam hal ini, penanda GK lalai ditandakan sebagai “kata nama” kepada setiap perkataan yang mempunyai satu makna atau perkataan yang tidak diketahui.

Ciri-ciri selanjutnya menggunakan modul dari Alfred et al. (2013) yang telah membina pendekatan berasaskan petua (RPOS) bagi korpus bahasa Melayu yang mengandungi petua bagi menanda GK bahasa Melayu yang mempunyai imbuhan

(awalan, apitan, akhiran), partikel *(-nya, -lah*) dan klitik *(-mu, -ku*). Berdasarkan RPOS, proses penandaan GK dimulakan dengan semakan ke atas kewujudan perkataan tersebut dalam kamus penandaan GK yang telah disediakan. Sekiranya perkataan tersebut wujud dan hanya mempunyai satu penanda sahaja, maka proses penandaan GK kata dianggap telah selesai. Sebaliknya, jika perkataan tersebut wujud dalam kamus tetapi mempunyai lebih dari satu kemungkinan penanda GK, jenis hubungan perkataan tersebut dikenalpasti untuk ditandakan dengan penanda GK yang sesuai. Sekiranya perkataan tersebut tidak wujud dalam kamus, perkataan tersebut akan diproses selaras dengan peraturan imbuhan sebelum ia diproses ke peringkat penandaan seterusnya. Senarai hubungan kata dan petua-petua imbuhan bagi mengenalpasti GK kata nama, kata kerja dan kata adjektif yang turut digunakan dalam kajian ini adalah seperti yang dipaparkan dalam Jadual 2.1 hingga Jadual 2.4.

Jadual 2.1 Senarai hubungan kata

Jenis GK	**Urutan Jenis GK yang Sah**
Kata Nama	Kata Sifat, Kata Adverba, Kata Kerja, Kata Nama, Kata Sendi,
Kata Kerja	Kata Bantu, Kata Adverba, Kata Nama, Kata Penekan, Kata Pembenda
Kata Adjektif (Sifat)	Kata Penguat, Kata Sendi
Kata Adverba	Kata Kerja, Kata Sendi, Kata Sifat, Kata Nama
Kata Arah	Kata Nama, Kata Sendi
Kata Sendi	Kata Nama, Kata Kerja, Kata Sifat
Kata Bantu Aspek	Kata Sifat, Kata Kerja, Kata Sendi
Kata Bilangan	Kata Nama
Kata Penekan	Kata Adverba, Kata Nama, Kata Hubung
Kata Pembenda	Kata Hubung, Kata Nama
Kata Hubung	Kata Nama, Kata Kerja, Kata Sendi, Kata Sifat
Kata Penguat	Kata Sifat
Kata Tanya	Kata Nama, Kata Kerja
Penanda Wacana	Kata Nama

Sumber: Alfred et al. (2013)

Jadual 2.2 Petua imbuhan kata nama

Petua	Imbuhan awal	Aksara seterusnya	Urutan aksara	Imbuhan Akhir	
				Boleh berakhir dengan	Perlu berakhir dengan
1a	pe	l, ng, ny, r, dan w	a-z	an	-
1b	pem	b dan p	a-z	an	-
1c	pen	c, d, j, sy dan z	a-z	an	-
1d	peng	g, h, k, kh, dan huruf vokal	a-z	an	-
1e	penge	-	a-z (3 – 4 aksara)	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,	-	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, isme, ik, is atau al.

Sumber: Alfred et al. (2013)

Jadual 2.3 Petua imbuhan kata adjektif

Petua	Imbuhan awal	Aksara seterusnya	Urutan aksara	Imbuhan Akhir	
				Boleh berakhir dengan	Perlu berakhir dengan
2a	ter, se, bi	-	a-z	-	-
2b	ke	-	a-z	an	-
2c	tidak bermula dengan di dan men	-	a-z	-	in, at, ah, iah, urutan huruf vokal, dan urutan huruf konsonan dan berakhir dengan i

Sumber: Alfred et al. (2013)

Jadual 2.4 Petua imbuhan kata kerja

Petua	Imbuhan awal	Aksara seterusnya	Urutan aksara	Imbuhan Akhir	
				Boleh berakhir dengan	Perlu berakhir dengan
3a	me	k, l, ng, ny, p,r, s,	a-z	-	-

bersambung…

…sambungan

		t, w dan y			
3b	mem	b, f, p dan v	a-z	kan atau i	-
3c	men	c, d, j, s, sy, t dan z	a-z	kan atau i	-
3d	meng	g, gh, h, k, kh, dan huruf vokal	a-z	-	-
3e	menge	-	a-z (3 – 4 aksara)	-	-
3f	memper	-	a-z	kan atau i	-
3g	ber	selain r	a-z	kan atau an	-
3h	bel	-	a-z	-	-
3i	Ter	selain r	a-z	-	
3j	Ke	-	a-z	-	an
3k	-	-	a-z	-	kan atau i
3l	di atau diper	-	a-z	kan atau i	-

Sumber : Alfred et al. (2013)

2.3.2 Pendekatan Berasaskan Statistik

Pendekatan berasaskan statistik menggunakan korpus latihan yang mengira kebarangkalian urutan yang diberikan oleh penanda yang boleh digunakan sebagai alternatif kepada pendekatan kekerapan perkataan (Kumawat & Jain 2015). Penandaan boleh ditentukan dengan mengetahui kebarangkalian bahawa perkataan tersebut akan berulang dengan *n* penanda sebelum ini, dimana nilai *n* disetkan kepada 1,2 atau 3 untuk tujuan praktikal. Algoritma Viterbi yang merupakan algoritma carian yang mengelak pengembangan polinomial dari carian pertama dengan mengurangkan (*trim*) pokok carian di setiap peringkat menggunakan anggaran kebolehjadian maksimum (MLE).

Kajian mengenai pendekatan berasaskan statistik bagi bahasa Melayu masih belum meluas dijalankan. Namun begitu, terdapat beberapa kajian lepas yang telah dijalankan oleh beberapa penyelidik seperti Hassan et al. (2011) yang menggunakan kaedah Model Markov Tersembunyi (MMT) dengan hasil ketepatan 67.9%, diikuti oleh Norshuhani et al. (2012) yang menggunakan kaedah n-gram dan fungsi persamaan *Dice Coefficient* untuk menyelaras tag dari bahasa Inggeris ke bahasa Melayu dengan hasil kejituan 86.87% dan Hassan et al. (2014) yang menjalankan kajian perbandingan antara kaedah MMT, Entropi Maksimum (EM) dan Mesin Vektor Sokongan (MVS) dengan hasil kejituan 99.23%.

Terdapat juga penyelidik yang membuat kajian penandaan GK bagi bahasa Indonesia seperti Pisceldo et al. (2009) yang menggunakan model kebarangkalian dengan kaedah *Conditional Random Fields* (CRF) dan EM dengan hasil ketepatan 97.57%, dan Wicaksono dan Purwarianti (2010) yang menggunakan model MMT dengan hasil ketepatan 96.50%,. Kajian bagi bahasa Indonesia ini turut diambilkira kerana bahasa ini menyerupai bahasa Melayu yang juga berasal dari filum bahasa Austris yang mana bahasa Austronesia merupakan salah satu daripada tiga rumpun utama daripada filum tersebut (Nik Safiah et. al 2015). Ini menunjukkan bahawa kajian mengenai penandaan GK bagi bahasa Melayu sedang giat dijalankan bagi mendapatkan keputusan yang terbaik.

2.3.3 Pendekatan Pembelajaran Mesin

Sementara pendekatan pembelajaran mesin (PM) pula merupakan satu set teknik statistik bagi mengenalpasti beberapa aspek teks seperti penandaan GK, pengecaman entiti dan sentimen analisis. Teknik tersebut dapat dinyatakan dalam model yang kemudiannya digunakan untuk teks lain (pembelajaran terselia) atau boleh menjadi satu set algoritma yang boleh digunakan dalam satu set data yang besar bagi mendapatkan maknanya (pembelajaran tidak terselia). Salah satu kelebihan pendekatan ini adalah model bahasa yang diperoleh mudah untuk disesuaikan dengan penandaan GK sedia ada dan kedua-dua modul tersebut boleh ditambahbaik dan dikembangkan secara berasingan tanpa bergantung terhadap satu sama lain (Màrquez 1999).

Beberapa kajian lepas yang dijalankan oleh penyelidik dengan menggunakan pendekatan ini adalah seperti Nakagawa et al. (2001) yang menggunakan teknik MVS bagi penandaan GK bahasa Inggeris dan jangkaan ketepatan bagi perkataan yang tidak diketahui dengan hasil ketepatan 97.1%. Gimpel et al. (2011) pula membina tagset, menganotasi data dan melakukan eksperimen ke atas data dalam media sosial *twitter* bahasa Inggeris menggunakan pendekatan PM terselia dengan hasil ketepatan 87.66% bagi 500 *tweets*. Bagi bahasa Arab, Habash dan Rambow (2005) menggunakan teknik penggabung majoriti (*majority combiner*) dan penggabung berasaskan keyakinan (*confidence-based combiner*) bagi melihat kesan penandaan GK ke atas dua sumber

berita yang berlainan dengan hasil ketepatan 98.1% bagi semua token. Sementara itu, Murata et al. (2002) pula membandingkan tiga kaedah yang menggunakan pendekatan PM iaitu senarai keputusan, EM dan MVS ke atas korpus dalam bahasa Thailand dengan hasil ketepatan terbaik diperoleh 96.1% melalui MVS.

2.3.4 Perbandingan Pendekatan Berasaskan Petua dan Statistik

Pendekatan berasaskan petua dan statistik mempunyai kelebihan dan kekurangan masing-masing. Jadual 2.5 memaparkan perbandingan antara kedua-dua pendekatan tersebut.

Jadual 2.5 Perbandingan pendekatan berasaskan petua dan statistik

Pendekatan	Berasaskan Petua	Berasaskan Statistik
Kelebihan	▪ Peraturan adalah tepat kerana ditulis berdasarkan teori linguistik ▪ Tidak memerlukan sumber pengkomputeran ▪ Mudah diperbaiki dan dianalisis dengan penambahan peraturan secara manual. ▪ Mudah melaksanakan analisis ralat	▪ Mudah dibina dan diselenggara (dengan kewujudan data) ▪ Tidak memerlukan kemahiran dan pengetahuan linguistik ▪ Tidak bergantung kepada masalah bahasa tempatan dan ketidakselarasan bahasa ▪ Dilatih dengan terjemahan manusia
Kekurangan	▪ Bagi teknik terselia, korpus perlu ditag terlebih dahulu secara manual ▪ Memerlukan kemahiran dan pengetahuan dalam bidang linguistik ▪ Ketidakselarasan bahasa jika ditanda secara manual oleh lebih daripada seorang.	▪ Memerlukan teks yang selari bagi mengelakkan penandaan yang tidak mengikut susunan tatabahasa ▪ Memerlukan sumber pengiraan pengkomputeran yang tinggi ▪ Sukar melaksanakan analisis ralat

Sumber: Costa-Jussà et al. (2012) dan Kumawat dan Jain (2015)

Berdasarkan Jadual 2.5, dapat disimpulkan bahawa kemahiran bahasa dan pengiraan pengkomputeran juga memainkan peranan yang penting dalam menentukan pendekatan yang dipilih selain daripada saiz data yang digunakan dan ketersediaan sumber data sedia ada.

2.4 PERKATAAN KABUR (*AMBIGUOUS*) DAN PERKATAAN YANG TIDAK DIKETAHUI (*UNKNOWN*)

Perkataan kabur adalah perkataan yang mempunyai lebih daripada satu makna atau dalam konteks GK, perkataan yang mempunyai lebih daripada satu penanda yang mungkin (Jurafsky & Martin 2016). Ini bermakna, terdapat unsur ketidakpastian dari segi leksikal dan struktur atau sintaksis, yang boleh mengandungi lebih daripada satu makna, bergantung kepada cara perkataan tersebut digunakan dalam ayat (Cao 2007). Kekaburan leksikal dianggap terjadi sekiranya terdapat kekaburan dalam satu perkataan (Hassan 2015). Oleh yang demikian, satu sistem PBT yang praktikal diperlukan bagi menyelesaikan masalah kekaburan dari segi kategori perkataan, struktur sintaksis dan skop semantik (Manning & Schiitze 1999).

Justeru, teknik nyahkabur makna-perkataan (NMP) diperkenalkan bagi mengenalpasti makna kata sebenar tentang perkataan dalam konteks tertentu (Navigli 2009) dan ia merupakan suatu tugasan yang sukar dan mencabar untuk melaksanakan dalam PBT. Pada asasnya, NMP banyak digunakan dalam aplikasi seperti capaian maklumat dan penterjemahan mesin keupayaan kerana pemahaman semantik masing-masing (Fulmari & Chandak 2014).

Sebagai penyelesaian terhadap masalah NMP ini, terdapat empat pendekatan yang boleh digunakan iaitu pembelajaran terselia, pembelajaran separa terselia, pembelajaran berasaskan pengetahuan dan pembelajaran tidak terselia (Pal & Saha 2015). Pembelajaran terselia menggunakan teknik PM daripada data yang telah dianotasi secara manual manakala pembelajaran separa terselia menggunakan data-data sama ada data yang telah dianotasi dan telah dikelaskan atau menggunakan data mentah (Fulmari & Chandak 2014). Dalam pembelajaran berasaskan pengetahuan pula, data-data untuk pengkelasan diambil dari sumber pengetahuan yang berbeza seperti kamus atas talian, *thesauri* atau *WordNet* (Fulmari & Chandak 2014). Pembelajaran tidak terselia pula tidak bergantung kepada sumber pengetahuan dari luar dan menggunakan data mentah (Fulmari & Chandak 2014).

NMP turut digunakan dalam proses penandaan GK untuk menyelesaikan masalah kekaburan. Sesetengah perkataan akan dinyahkabur kepada tahap homograf