PENGUKURAN KESETARAAN AYAT BAGI PENILAIAN ESEI PENDEK BAHASA MELAYU BERPANDUKAN KAEDAH PERANAN TEMATIK DAN RANGKAIAN SEMANTIK

MOHD AZWAN BIN MOHAMAD@HAMZA

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

PENGAKUAN

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

20 November 2019

MOHD AZWAN BIN MOHAMAD@HAMZA
P59305

PENGHARGAAN

Assalamualaikum W. B. T...

Segala puji bagi Allah S.W.T, Tuhan semesta alam. Salawat dan salam ke atas junjungan besar Nabi Muhammad S.A.W, keluarga dan para sahabat baginda serta kaum Muslimin dan Muslimat.

Bersyukur saya ke hadrat Allah S.W.T kerana atas limpah kurniaNya serta keizinanNya, dapatlah jua saya menyiapkan kajian saya ini.

Di kesempatan ini juga saya ingin merakamkan jutaan terima kasih dan penghargaan ikhlas buat Profesor Dr. Mohd Juzaiddin Ab Aziz selaku penyelia dan PM Dr. Nazlia Omar selaku penyelia bersama, atas bimbingan dan dorongan yang diberikan di sepanjang tempoh kajian ini dibuat.

Penghargaan juga turut ditujukan kepada semua yang terlibat samada secara langsung atau tidak langsung dalam membantu menjayakan kajian ini.

ABSTRAK

Penilaian Esei Pendek Bersepadu (PEPB) merupakan satu bentuk penilaian subjektif yang menitikberatkan isi-isi penting berbanding gaya penulisan menggunakan samada kaedah-kaedah statistik atau linguistik atau juga gabungan antara keduanya. Pada peringkat awal kajian, PEPB hanya mampu mengukur kualiti esei pendek berdasarkan beberapa parameter asas seperti purata panjang kata, pembahagian panjang ayat, jenis nahu dalam ayat, peratusan kata kerja pasif dan peratusan kata nama. Walaubagaimanapun, kemajuan bidang pemprosesan bahasa tabii dan capaian maklumat telah merintis pada penghasilan rekabentuk PEPB yang lebih berkesan untuk mengukur kesetaraan ayat antara skema jawapan dan jawapan pelajar. Sementara kebanyakan penyelidik telah mengaplikasikan pengetahuan semantik untuk mengukur makna perkataan yang spesifik bagi Bahasa Inggeris, kekurangan pangkalan data semantik leksikal Bahasa Melayu telah menyebabkan timbulnya kekangan untuk menyelesaikan masalah kekaburan dan penjejakan mengundur dalam padanan nahu. Pengetahuan semantik mengumpulkan kata nama, kata kerja, kata adjektif dan kata sifat dalam kumpulan *synset*. Kedua-dua *synset* seterusnya akan dipadankan berdasarkan konsep semantik dan hubungan leksikal berbanding kamus. Namun begitu, pengukuran kesetaraan ayat perlu mempertimbangkan keseluruhan struktur ayat terlebih dahulu. Maklumat statistik, susunan perkataan, peraturan nahu dan struktur sintaksis merupakan kaedah-kaedah yang digunakan kajian terdahulu yang menitikberatkan struktur subjek dan predikat dalam setiap ayat. Tetapi, kaedah tersebut memerlukan saiz korpus yang besar, kepelbagaian padanan susunan, peraturan yang kompleks dan kewujudan imbuhan yang memberi kesan ke atas stuktur sintaksis yang tidak tepat. Dengan itu, kajian ini cuba menyelesaikan masalah tersebut ke atas Bahasa Melayu dengan menggunakan Teknik Rangkaian Semantik berdasarkan pengukuran *Wu & Palmer*, yang mana ianya mengira kadar kerelatifan dua *synset* dengan mengambilkira kedalaman semantik pada dua peringkat; penandaan leksikal dan pengiraan kesetaraan ayat. Namun begitu, sebelum pengukuran kesetaraan ayat dibuat, token kata dan frasa yang telah ditanda akan distrukturkan menurut Petua Peranan Tematik yang terdiri daripada Pelaku, Penderita, Tema, Tempat, Pemanfaat, Pengalami, Sumber, Bilangan dan Masa supaya pemadanan token yang signifikan akan dikelaskan mengikut hubungan tema masing-masing dalam ayat. Kajian ini diuji ke atas set data kursus Pengkompil dalam Bahasa Melayu yang terdiri daripada 13 soalan dan 185 jawapan pelajar merangkumi soalan pasif, ringkas, negatif, gabungan (lebih dari satu objek) dan kompleks (lebih dari satu objek, subjek dan kata kerja). Dengan mengaplikasikan kedua-dua teknik ini dalam metodologi spesifik, hasil yang diperolehi cukup meyakinkan. Bagi membuktikan keberkesanan teknik-teknik tersebut, hasil ujian menggunakan Teknik *Pola Grammar* dijadikan penanda aras pengujian dapatan, kejituan dan ukuran-f masing-masing adalah 95.83%, 91.33% dan 93.53% menggunakan Petua Peranan Tematik dan Teknik Rangkaian Semantik berbanding 83.39%, 81.36% dan 82.36% menggunakan Teknik *Pola Grammar*.

SENTENCE SIMILARITY MEASUREMENT ON SHORT MALAY ESSAY ASESSMENT BASED ON THEMATIC ROLES AND SEMANTIC NETWORK

ABSTRACT

An Intergrated Short Essay Assessment (ISEA) system is a computer program that evaluates the subjective answer by emphasizing on the main points rather than the writing style using either statistical, linguistic or combination of both methods. The critical factor that determines the success of an ISEA system is the grade of an essay generated by the system should be acceptable and similar to human's assessment, relatively. At this preliminary stage, ISEA is only capable to consider fundamental assessment parameters such as the average word length, the distributions of the sentence length, grammatical types of sentences, the percentage of passive voice verbs and the percentage of nouns. However, the advancement in the natural language processing and information retrieval field have paved the creation of more effective design of ISEA to measure the similarities between the answer scheme and the students' answer. While most researchers employ the semantic knowledge in order to measure the specific sense of words for English language, the lack of Malay lexical semantic database has lead to the constraint on solving ambiguity and backtracking in grammar matching. Semantic knowledge groups the nouns, verbs, adjectives and adverbs into a group of synset. Both synsets are then matched based on the conceptual semantic and lexical relations rather than a dictionary. Nevertheless, measuring a sentence similarity requires consideration of the whole sentence structure. Statistical information, word order, grammar rules and syntactic structure were the methods used in previous research, which take into account the context of words in every sentence. But, these methods require huge size of corpus, varieties order matching, complicated rules and existing of stem that will reflect to inaccurate syntactic structure. Hence, this study attempts to solve such problems on Malay language using Semantic Network Technique based on Wu & Palmer measurement, which calculates the relatedness by considering the semantic depths of two synsets at two levels; lexical tagging and sentence similarity calculation. However, tagged word and phrase tokens will be labelled according to the rules of Thematic Role Rules that consists of Actor, Patient, Theme, Place, Beneficier, Experiencer, Source, Amount and Time. Those significant tokens is classified based on the specific thematic relations in a particular sentence, at prior. This study has been tested on the data set of Compiler course in Malay language that consists of 13 questions and 185 students' answer, which comprise of passive, simple, negative, conjoined (more than one objects) and complex (more than one object, subject and verb). By employing both of the techniques in a specific methodology, the results has been promising. The result of the Pola Grammar Technique is set as a benchmark to prove the effectiveness of the techniques. The recall, precision and f-measure test gain 95.83%, 91.33% and 93.53% using Thematic Role Rules and Semantic Network Technique compared to 83.39%, 81.36% and 82.36% using Pola Grammar Technique, respectively.

KANDUNGAN

SENARAI JADUAL

SENARAI RAJAH

SENARAI SINGKATAN

Penggunaan dalam Bahasa Inggeris		Penggunaan dalam Bahasa Melayu	
Adj	Adjective	Adj	Adjektif
AI	Artificial Intelligent	KP	Kepintaran Buatan
AP	Adjective Phrase	FA	Frasa Adjektif
CNN	Convolutional Neural Network	RNK	Rangkaian Neural Konvolusi
CP	Compliment Phrase	FP	Frasa Pelengkap
FP	False Positive	PS	Positif Palsu
GR	Grammatical Relationship	HN	Hubungan Nahu
IEA	Integrated Essay Assessment	PEB	Penilaian Esei Bersepadu
IR	Information Retrieval	DM	Dapatan Semula Maklumat
ISEA	Integrated Short Essay Assessment	PEPB	Penilaian Esei Pendek Bersepadu
LCS	Least Common Subsummer	HST	Hasiltambah-Sub Sepunya Terkecil
MLR	Multiple Linear Regression	PLB	Pengunduran Linear Berganda
MRD	Machine-Readability Dictionary	KKM	Kamus Kebolehbacaan Mesin
N	Noun	KN	Kata Nama
NLP	Natural Language Processing	PBT	Pemprosesan Bahasa Tabii
NP	Noun Phrase	FN	Frasa Nama
NPP	Noun Preposition Phrase	FS	Frasa Sendi Nama

NTV	Non-Transitive Verb	KKTT	Kata Kerja Tak Transitif
PG	Pola Grammar	PN	Pola Nahu
POS	Part-of-Speech	GK	Golongan-Kata
PPV	Positive Predictive Value	NJP	Nilai Jangkaan Positif
QW	Question Word	KT	Kata Tanya
SLR	Stepwise Linear Regression	PLT	Pengunduran Linear Berturutan
SN	Semantic Network	RS	Rangkaian Semantik
SP	Subject-Predicate	SP	Subjek-Predikat
SVD	Single Value Decomposition	PNT	Penghuraian Nilai Tunggal
SVM	Space Vector Model	SVM	Model Ruang Vektor
TFIDF	Term Frequency–Inverse Document Frequency	KIKDP	Kekerapan Istilah Kekerapan Dokumen Pembalikan
TP	True Positive	PP	Positif Pasti
TPR	True Positive Rate	TPP	Kadar Positif Pasti
TR	Thematic Roles	PT	Peranan Tematik
TV	Transitive Verb	KKT	Kata Kerja Transitif
V	Verb	KK	Kata Kerja
VP	Verb Phrase	FK	Frasa Kerja
WSB	Word Sense Disambiguation	PMP	Penyahtaksaan Makna Kata

SENARAI ISTILAH

Synset	Token signifikan dalam ayat.
Analitikal	Mampu atau berkebolehan untuk menganalisis sesuatu.
Antonim	Nama berlawanan.
Argumen	Token frasa nama subjek dan objek yang dipengaruhi oleh kata kerja.
Bahasa tabii	Bahasa yg digunakan oleh manusia dlm keadaan semula jadi.
Bilangan	Bilangan di mana perlakuan atau keadaan yang dilahirlan oleh predikat.
Fitur	Ciri-ciri.
Hipernim	Nama super-kelas.
Hiponim	Nama sub-kelas.
Holonim	Nama keseluruhan.
Kata-ke-kata	Perbandingan antara dua kata / frasa (token) dalam dua ayat.
Masa	Masa di mana perlakuan atau keadaan yang dilahirlan oleh predikat.
Meronim	Nama sebahagian.
Pelaku	Suatu yang secara sedar memulakan perlakuan yang dilahirkan oleh predikat.
Pemanfaat	Entiti yang memperolehi sesuatu keadaan yang dilahirkan oleh predikat.
Penderita	Orang atau benda yang mengalami perlakuan yang

	dilahirkan oleh predikat
Pengalami	Entiti yang mengalami sesuatu keadaan yang dilahirkan oleh predikat.
Penilaian Formatif	Penilaian yang berterusan dalam proses pengajaran dan pembelajaran.
Penilaian Sumatif	Penilaian untuk membezakan antara peringkat pencapaian pelajar.
Semantik	Kajian tentang makna per-kataan dan penambahan makna sesuatu kata.
Sinonim	Nama yang hampir sama.
Sumber	Entiti dari mana sesuatu diubah hasil daripada perlakuan yang dilahirlan oleh predikat.
Tema	Orang atau benda yang mengalami perlakuan yang dilahirkan oleh predikat.
Tempat	Tempat di mana perlakuan atau keadaan yang dilahirlan oleh predikat.
Troponim	Nama sifat.

BAB I

PENGENALAN

1.1 PENDAHULUAN

Dalam dunia pendidikan, penilaian esei merupakan satu tugas yang penting untuk menilai kualiti penulisan esei mengenai topik tertentu. Menurut Biggs (1998), penilaian esei ditakrifkan sebagai tindak balas berterusan ke atas soalan tertentu, di mana pelajar itu telah menerima arahan (soalan), disediakan dalam tempoh masa yang pelajar itu sendiri, dan tertakluk kepada beberapa jenis penilaian sumatif. Proses menulis tugasan berbentuk esei melibatkan analitikal, pemikiran kritikal dan kemahiran komunikasi yang mana ianya dicadangkan untuk pendekatan pembelajaran mendalam; yang membolehkan pelajar memilih, menyusun-atur dan menerangkan pengetahuan dan kefahaman mereka dengan cara membentangkan pengetahuan teori meta dan bersyarat jika mereka memilikinya (Blood 2011; Boulton-Lewis 1995).

Soalan penilaian esei boleh dibahagikan kepada dua jenis; esei panjang dan esei pendek. Esei panjang merupakan esei teks bebas di mana para pelajar diberi sebuah topik untuk dibincangkan dalam bentuk esei panjang (Mohd Juzaiddin et. al 2009). Penilaian bagi esei panjang biasanya terdiri daripada nahu, penggunaan, mekanik (ejaan, tanda bacaan, huruf besar dan perenggan) dan gaya (Attali & Burstein 2006). Manakala, esei pendek ditulis dalam bentuk ayat-ayat ringkas di mana gaya penulisan tidak dititiberatkan untuk pemarkahan (Mohd Juzaiddin et. al 2009). Memandangkan bilangan kata yang wujud dalam esei pendek adalah terhad, maka setiap kata atau frasa dalam ayat adalah signifikan untuk menyumbang sebahagian markah pada penilaian esei tersebut. Namun, penilaian esei secara manual akan memakan masa yang banyak, melibatkan kos yang tinggi dan berkemungkinan untuk

dipengaruhi faktor emosi penilai (Ramlingam et. al 2018; Zhang 2014). Oleh itu, kajian dalam bidang Penilaian Esei Bersepadu (PEB) dilihat sebagai penyelesai masalah tersebut.

1.2 LATAR BELAKANG KAJIAN

PEB didefinisikan sebagai teknologi komputer yang mempunyai kemampuan untuk menilai kualiti sebenar sesebuah penulisan (Shermis & Burstein 2016). PEB mula diperkenalkan pada kira-kira lima dekad yang lalu, pada tahun 1966 oleh Ellis Page (Page 1994). Walaubagaimanapun, sebahagian pengkaji dalam bidang Pemprosesan Bahasa Tabii (PBT) mempercayai bahawa PEB telah pun mula diperkenalkan sekitar awal tahun 1960-an. Sungguhpun begitu, jenis fitur-fitur yang mampu diekstrak daripada teks secara automasi hanya terhad pada fitur-fitur luaran (Hearst 2000). Dari situ, timbulnya perbahasan antara para pengkaji dari segi persoalan samada PEB mampu untuk menggantikan penilai manusia dalam menilai kualiti penulisan esei. PEB yang baik seharusnya menunjukkan kolerasi yang kuat dari segi tahap penilaian antara sistem dan manusia.

Walaupun demikian, menurut Yang et. al (2002), penggunaan sistem PEB berisiko menyebabkan berlakunya terlebih-kebergantungan terhadap penilaian fitur-fitur luaran, kurang pertimbangan ke atas isi kandungan dan kreativiti, serta membuka ruang untuk terjadinya beberapa bentuk penipuan cara baru dalam menjawab soalan. Dalam kata lain, PEB dilihat hanya menilai penulisan berdasarkan fitur-fitur yang tidak-langsung menggambarkan kualiti sebenar penulisan. Oleh itu, ianya boleh dipersoalkan samada PEB mampu menilai esei secara efektif jika, “esei tersebut tersusun dari segi susunan isi dan ayat yang baik, tetapi dengan mekanik yang lemah atau nahu yang baik, tetapi dengan kesalahan ejaan yang banyak” (Calfee 2000).

Walaubagaimanapun, beberapa faktor yang telah mendorong para penyelidik untuk meneruskan kajian dalam bidang ini: [1] praktikal; penilaian esei melibatkan kos yang tinggi dan memakan masa yang banyak; [2] konsisten; penilaian esei bersifat subjektif, oleh itu sukar untuk membuat penilaian yang sangat konsisten; [3] maklumbalas; mampu memberi penerangan yang terperinci bagi setiap markah yang

diperolehi dan kesalahan yang dikenalpasti (Ramalingam et. al. 2018; Zhang 2014; Lonsdale & Strong-Krause 2003) and [4] umum; tidak terikat pada mana-mana domain (Dikli 2006).

Tumpuan utama bagi sistem PEB adalah markah yang dijana oleh sistem ini seharusnya boleh diterima iaitu menghampiri tahap penilaian manusia (Mohd Juzaiddin et. al 2009). Menurut Sakaguchi et. al (2015), secara umumnya, kebanyakan kajian dalam penilaian esei pendek menggunakan salah satu daripada dua kaedah tersebut:

i. pendekatan berpandukan-maklumbalas yang menggunakan fitur-fitur terperinci yang diekstrak daripada maklumbalas pelajar itu sendiri (n-gram, dan sebagainya) dan mempelajari fungsi penilaian melalui maklumbalas penilai-manusia.

ii. pendekatan berpandukan-rujukan yang membandingkan jawapan pelajar dengan skema jawapan (kesetaraan ayat).

Namun begitu, masih ramai pengkaji menggunakan kaedah berpandukan kesetaraan ayat untuk membuat penilaian esei pendek (Pawar & Mago 2018; Uswatun Hasanah et. al 2019; Pulman dan Sukkarieh 2005; Cutrone dan Chang 2011; Ali Mutfah & Mohd Juzaiddin 2013). Mereka membuat penilaian esei pendek dengan membandingkan kesetaraan antara dokumen skema jawapan dan dokumen jawapan pelajar (Suzen et. Al 2018). Dengan membandingkan jawapan pelajar dan skema jawapan, penilaian dibuat berdasarkan tahap kesetaraan antara kedua-dua dokumen tersebut yang terdiri daripada satu atau lebih dari satu ayat.

Pendekatan pengukuran kesetaraan ayat perlu mengenalpasti kaedah yang berkesan dalam bidang dan aplikasi PBT (Li et. al 2006). Penilaian teks atau ayat pendek tidak seharusnya bergantung kepada kekerapan kewujudan bilangan kata berbanding teks panjang.

Menurut Li et. al (2006) lagi, kemampuan bahasa tabii yang bersifat fleksibel adalah sebab utama yang memungkinkan kita membina ayat pendek yang berbeza tapi

mempunyai makna yang sama. Beberapa kajian dalam bidang PBT telah dibuat untuk tujuan ini. Karov & Edelman (1998) memperkenalkan Kaedah Penyahtaksaan Makna Kata (*Word Sense Disambiguation)* (KPM) yang berpandukan-kesetaraan menggunakan teks korpus dan Kamus Boleh Dibaca-Mesin (*Machine-Readability Dictionary*) (KBD). KBD diimplementasi menggunakan teori atribut (Karov & Edelman 1998; Mohamad Albared et. al 2009; Mohamad Albared et. al 2010; Mohd Juzaiddin et. al 2006), manakala bagi kaedah berpandukan-korpus, ianya diimplementasikan menggunakan rangkaian semantik dan statistik (Li et. al 2006). Dalam kajian ini, teknik yang dikenalpasti mampu membuat penilaian secara holistik untuk penilaian peperiksaan esei pendek akan menjadi objektif utama. Perkara yang menjadi fokus utama kajian adalah jika kaedah yang dipilih tidak mampu untuk mengukur kesetaraan ayat dengan tepat dan boleh-dipercayai berdasarkan suatu piawaian yang ditetapkan, maka kaedah tersebut tidak akan diterima.

Algoritma pengukuran kesetaraan ayat bagi penilaian esei pendek melibatkan set jawapan pelajar dan skema pemarkahan semakin banyak digunakan dalam kajian terkini (Pawar & Mago 2018, Ho et. al. 2010, Manna & Gedeon 2010, Yi & Qiang 2009). Kedua-dua dokumen teks tersebut dibuat perbandingan untuk mendapatkan markah yang tepat dan boleh-diterima berdasarkan kaedah kesetaraan ayat (Dikli 2006; Erikson 2000; Fitzgerald 1994; Mohd Juzaiddin 2008). Penilaian berdasarkan kesetaraan ayat menjadi kompleks apabila ia melibatkan perbandingan lebih dari satu struktur ayat. Tahap kompleks ayat tersebut juga dipengaruhi oleh beberapa faktor iaitu kedudukan kata dalam struktur ayat dan kekaburan kata dalam ayat (Lee et. al 2014; Pearce 2015; Kadupitiya et. al 2016; Pawar & Mago 2018).

1.3 PENYATAAN MASALAH

Kesetaraan ayat tidak boleh diukur dengan hanya membuat perbandingan secara langsung memandangkan ianya dibina menggunakan kata, frasa dan struktur yang berbeza (Shahrul Azman et. al 2007; Mohd Juzaiddin et. al 2008; Ho et. al 2010a; Ho et. al , 2010b). Li et. al (2006) menyatakan bahawa terdapat dua kelemahan yang wujud dalam kesetaraan ayat apabila kita menggunakan kaedah pemadanan kata-ke-kata (*word-to-word*) iaitu ianya memerlukan penglibatan manusia secara intenstif dan

ianya juga tidak boleh diaplikasikan dengan mudah pada domain yang lain. Dengan itu, kita tidak boleh mengukur kesetaraan ayat dengan hanya berdasarkan kepada perbandingan makna langsung antara dua kata. Selain itu, kedudukan kata yang mempengaruhi subjek dan predikat dalam ayat turut mempengaruhi pengukuran kesetaraan ayat yang dibuat.

Bagi mendapatkan kesetaraan hubungan sematik dan susunan kata, pendekatan semantik berpandukan-vektor beserta dengan maklumat kandungan leksikal dan korpus telah diaplikasikan. Shahrul Azman et. al (2007) menggunakan rangkakerja yang hampir sama untuk mengukur kesetaraan sematik dan susunan kata tersebut. Walaubagaimanapun, evolusi kesetaraan kata-ke-kata mempunyai beberapa isu iaitu perwakilan ayat adalah kurang efisyen disebabkan oleh dimensi vektor yang besar (Li et. al 2006). Saiz dimensi vektor adalah besar berbanding bilangan kata dalam ayat dan memerlukan kumpulan kata untuk domain spesifik bagi menjamin ketepatan pengukuran kesetaraan ayat yang lebih baik. Selain itu, antara isu lain ialah ayat yang mempunyai makna yang sama tidak semestinya dibina dari sejumlah kata yang sama.

Kajian penyelidik terdahulu menunjukkan struktur ayat memainkan peranan penting pada peringkat awal dalam mengukur kesetaraan ayat (Kuboň et. al 2013, Li et. al 2006, Mohd Juzaiddin 2008). Ayat aktif dan pasif sebagai contoh, mempunyai struktur subjek dan predikat yang berbeza, namun setara jika diaplikasikan menggunakan kaedah yang betul (Li & Li 2015). Li et. al (2006) telah menggunakan kaedah maklumat statistik bagi menentukur struktur ayat, namun ianya memerlukan saiz korpus yang besar. Manakala Ho et. al (2010) dan Pawar & Mago (2018) menggunakan kaedah susunan kata, namun disebabkan terdapatnya kepelbagaian padanan susunan, kaedah ini mengalami kesukaran untuk diaplikasikan. Mia & Ayu (2011) pula menggunakan kaedah petua bahasa, namun petua yang kompleks menyukarkan padanan struktur ayat dibuat. Bagi Wafa Wali et. al (2017) dan Ma & Suel (2016) yang menggunakan kaedah struktur sintaksis, mendapati apabila dua ayat yang mempunyai struktur sintaksis yang sama (subjek + kata kerja + objek) tetapi kelas semantik objek-objek tersebut berbeza, maka pasangan ayat tersebut dikira setara secara sintaksisnya, sedangkan sebenarnya kedua-dua ayat tersebut adalah langsung tidak setara menurut penilai manusia. Kesetaraan tematik mengambilkira

corak bagi konsep (argumen) dan hubungan yang wujud antara keduanya (Khan & Mustafa 2014). Berdasarkan kajian yang dibuat, kesetaraan tematik telah digunakan untuk capaian maklumat dan berjaya mengenalpasti konteks bagi konsep tersebut. Dor et. al (2018) juga telah mewakilkan konteks bagi kata kunci melalui corak kata kunci untuk carian semantik yang berkesan menggunakan kesetaraan tematik.

Isu seterusnya melibatkan padanan kata dan frasa yang signifikan dalam ayat. Bagi kajian yang melibatkan Bahasa Inggeris, masalah ini sudah tidak lagi ketara. Ini kerana kebanyakan penyelidik menggunakan Kaedah Rangkaian Semantik untuk mengukur tahap kesetaraan hubungan semantik antara dua *synset* (argumen) (Pawar & Mago 2018, Li 2006). Sebaliknya bagi Bahasa Melayu, kegagalan kaedah yang digunakan untuk menafsir pengetahuan semantik bagi mengekstrak makna spesifik setiap frasa menyebabkan wujudnya masalah kekaburan dan penjejakan mengundur dalam padanan nahu.

1.4 MATLAMAT DAN OBJEKTIF KAJIAN

Kajian ini bertujuan untuk membuat penilaian ke atas esei pendek Bahasa Melayu dengan membandingkan kesetaraan antara dokumen (skema jawapan dan jawapan pelajar) menggunakan pendekatan pengukuran kesetaraan ayat. Bagi melaksanakan tujuan ini, objektif kajian ini adalah untuk:

i. Mengenalpasti faktor-faktor penentu seperti jenis kata kerja, imbuhan yang disisipkan ke atas kata kerja dan kewujudan kata khas dalam pembangunan set Petua Peranan Tematik beserta proses penormalan ayat pada peringkat latihan merangkumi peranan Pelaku, Penderita, Tema, Tempat, Pemanfaat, Pengalami, Masa dan Bilangan.

ii. Membangunkan algoritma untuk menanda peranan tematik ke atas setiap argumen subjek, kata kerja, objek dan argumen-argumen signifikan lain yang wujud berpandukan kepada set Petua Peranan Tematik yang telah dibina pada peringkat ujian dan memadankan

peranan-peranan tersebut antara set jawapan pelajar dan skema jawapan dengan mempertimbangkan konteks argumen dalam ayat.

iii. Mengaplikasikan kaedah pengukuran Wu & Palmer (wup) dalam rangkaian semantik Bahasa Melayu berasaskan senibina *Wordnet* bagi mengukur kesetaraan hubungan semantik pada peringkat *synset* dan ayat berpandukan kepada peranan tematik yang telah ditanda.

1.5 SIGNIFIKAN KAJIAN

Kajian ini akan menyumbang kepada pengetahuan baru dalam bidang pendidikan dalam skop linguistik menggunakan pendekatan sains komputer. Signifikan kajian ini adalah:

i. Proses pernormalan ayat yang mengambilkira kewujudan subjek dan predikat, faktor imbuhan dalam kata dan frasa, serta kata ganti nama, boleh digunakan dalam kajian domain liguistik yang lain.

ii. Petua peranan tematik pada aras ayat dijangka mampu untuk meningkatkan tahap kebolehpercayaan dan ketepatan penilaian esei pendek. Ini kerana penggunaan kaedah lingustik dilihat mampu membuat penilaian yang lebih baik menghampiri kadar penilaian oleh pakar manusia.

1.6 SKOP KAJIAN

Skop kajian lebih berkisar kepada set data. Memandangkan kajian ini melibatkan penilaian esei pendek berautomasi, skop adalah:

i. Bilangan ayat maksima bagi setiap jawapan pelajar tidak lebih daripada 15 ayat.

ii. Dokumen (set skema jawapan dan jawapan pelajar) ditulis dalam Bahasa Melayu sahaja.

iii. Set data adalah terdiri daripada 13 set soalan dan set skema jawapan beserta 185 set jawapan pelajar berdasarkan ujian subjek Pengkompil yang merangkumi ayat-ayat ringkas dan kompleks (majmuk gabungan, pancangan dan campuran).

iv. Semua dokumen adalah dalam format esei salinan lembut, bukan tulisan tangan.

v. Algoritma yang dibangunkan akan memproses bahasa sehingga aras semantik dalam bidang pemahaman bahasa tabii.

1.7 RUMUSAN

Bab ini membincangkan perkara pokok iaitu masalah yang wujud dalam kajian semasa bagi domain berkaitan serta tujuan kajian ini dibuat. Secara tidak langsung ianya menjadi motivasi untuk memungkinkan penyelidikan ini diteruskan. Turut dinyatakan sebahagian hipotesis dari segi kaedah-kaedah yang akan digunakan dengan harapan dapat menyelesaikan masalah semasa tersebut. Dari situ, sumbangan kajian ini dalam bidang berkaitan diperjelaskan. Namun begitu, bagi membolehkan kajian ini disahkan secara ilmiah, bab ini juga telah menyatakan ruang lingkup atau skop kajian supaya hasil akhir kajian tidak boleh dipertikaikan.

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

Esei dianggap oleh ramai penyelidik sebagai satu alat yang sangat sesuai untuk menilai hasil pembelajaran, di mana ianya mampu mengukur kebolehan seseorang untuk mengingat, menyusun dan menggabungkan idea, kemampuan untuk menerangkan sesuatu dalam bentuk penulisan dan juga kemampuan untuk mengenalpasti, berbanding hanya menafsir dan mengaplikasi data (Valenti et. al 2003). Walaupun isu kesahihan peralatan Penilaian Esei Bersepadu (PEB) masih lagi diperdebatkan, namun begitu beberapa PEB telah berjaya dihasilkan dan digunakan. Dengan menggunakan teknologi Pemprosesan Bahasa Tabii (PBT) dan teknik-teknik statistik yang terkini untuk menganalisa gaya dan kandungan penulisan esei, ianya mampu mencapai lebih daripada 0.90 kolerasi markah komputer-manusia.

Pada tahun 1966, Ellis Page, perintis bagi PEB telah memulakan kajian dengan merekabentuk sebuah program komputer dinamakan *Project Essay Grade* (PEG) (Wang & Brown 2007). Dengan kekangan kemampuan statistik yang terhad pada masa itu, PEG cuba untuk menentukan kombinasi optimum pemberat fitur untuk membuat peramalan yang terbaik, iaitu menghampiri penilaian guru menggunakan Teknik Pengunduran Linear Berganda (PLB). Dengan kata lain PEG menekankan penilaian esei berpandukan kualiti penulisan asas tanpa mengambilkira kandungan. Rekabentuk pendekatan PEG adalah berasaskan kepada *proxes* yang terdiri daripada panjang esei, bilangan kata depan, kata ganti relatif dan penggolongan kata, sebagai pengukur kepada tahap kerumitan sesuatu struktur ayat (Valenti et. al 2003).

Sekitar tahun 1980-an, corak penilaian telah berubah daripada hanya menilai esei kepada memberi maklumbalas kepada pelajar dan guru. Sistem *Writer's Workbench* (WWB) yang dibangunkan oleh *American Telephone & Telegraph* (AT&T) telah direkabentuk untuk memberi maklumbalas kepada penulis dari segi "ejaan, sebutan dan kebolehbacaan" (Hearst 2000). WWB menilai beberapa fitur berkaitan dengan gaya, purata panjang kata, pembahagian panjang ayat, jenis nahu dalam ayat, peratusan kata kerja pasif dan peratusan kata nama yang telah melalui proses pernormalan (Burstein & Wolska 2003). Maklumbalas yang diperolehi daripada sistem ini boleh digunakan untuk meningkatkan skil penulisan pelajar pada masa akan datang.

Perkembangan PBT dan Dapatan Semula Maklumat (DSM) pada lewat 1990-an telah membuka satu lagi lembaran baru dengan penghasilan lebih banyak rekabentuk PEB yang lebih maju. Pada masa itu, sebanyak tiga PEB utama telah berjaya dibangunkan; *e-rater*, *Intelligent Essay Assessor* (IEA) dan *IntelliMetric*. Dengan menggaplikasikan Pengunduran Linear Berturutan (PLB) untuk menentukan model penilaian yang mampu membuat peramalan terbaik menghampiri tahap penilaian manusia, *e-rater* menilai esei berpandukan kepelbagaian sintaksis, kandungan topik dan susunatur idea (Attali & Burstein 2006).

Dengan menggunakan pendekatan *Latent Semantic Analysis* (LSA), Landauer & Laham (2000) telah membangunkan IEA yang mana menekankan penyediaan maklumbalas yang mengambilikira tiga elemen iaitu isi kandungan, gaya dan mekanik. Seterusnya, para penyelidik *Vantage Learning* melaporkan bahawa mereka telah mengadaptasikan Kepintaran Buatan (KP) bersama-sama dengan PBT dan teknologi statistik dalam membangunkan Intellimetric dan dengan ini, ianya mampu untuk menganalisa lebih dari 300 fitur-fitur pada aras semantik, sintaksis dan hujahan (Valenti et. al 2003).

Dalam kerancakan para penyelidik PEB bertungkus-lumus mengekstrak fitur-fitur paling signifikan dalam penilaian penulisan esei panjang, sebaliknya sebahagian dari mereka mula mengkaji Penilaian Esei Pendek Bersepadu (PEPB). Kebanyakan teknik yang biasanya sesuai diaplikasikan ke atas esei panjang mungkin tidak lagi

sesuai diimplementasikan ke atas esei pendek. Ini kerana esei panjang biasanya mengandungi kadar kekerapan kewujudan kata yang tinggi (Li et. al 2006). Dalam esei pendek, setiap kata yang wujud mempunyai kebarangkalian yang tinggi untuk bersifat signifikan dalam menyumbang kepada markah tertentu. Sistem pemarkahan esei pendek direkabentuk untuk jawapan yang ringkas dan berbentuk fakta di mana kriteria betul atau salah adalah jelas (Raheel Siddiqi 2010). Markah yang diberi adalah lebih berdasarkan isi kandungan berbanding gaya penulisan.

Kini terdapat beberapa buah sistem yang telah berjaya dibangunkan sesuai untuk penilaian esei pendek berautomasi. Sistem *Oxford-UCLESS* menggunakan pendekatan corak iaitu ianya dimulakan dengan set kata dan sinonim seterusnya membuat carian dalam tetingkap teks (esei) untuk menghasilkan corak yang baru (Mohler & Mihalcea 2009). Kaedah DSM telah diadaptasikan untuk digunakan dalam sistem ini yang memproses ayat-ayat tak-gramatis dan tak-lengkap dalam kebanyakan kes peperiksan *UCLES*. *C-rater* pula merupakan enjin penilaian esei pendek, dibangunkan oleh *ETS Technologies*, yang mana telah direkabentuk untuk memberikan maklumbalas jawapan yang diperolehi bagi soalan yang dinilai. Sistem ini mampu mengukur kefahaman isi kandungan ayat. Ianya menggunakan struktur frasa predikat rujukan pronominal, analisa morfologi dan sinonim bagi menilai keseluruhan atau sebahagian soalan esei pendek (Leacock & Chodorow 2003). Selain itu, *Automark*, yang dibangunkan untuk penilaian berautomasi yang robust ke atas teks-bebas pendek. Kaedah DSM digunakan untuk mengekstrak konsep atau makna di sebalik teks-bebas tersebut dan pembangun telah berusaha sebaik mungkin untuk memastikan sistem ini mempu bertoleransi dengan baik dari segi kesilapan menaip, mengeja, sintak dan sebagainya (Raheel Siddiqi 2010). Ianya memberi penekanan ke atas analisa kandungan tanpa mengabaikan fitur gaya penulisan.

2.2 KAEDAH PENILAIAN PEPB

Kebanyakan kaedah PBT kadangkala mengimplementasi pemprosesan yang asas pada peringkat awal seperti pembetulan ejaan, pembetulan tanda baca, pembetulan kata ganti nama, *lemmatization* dan penanda (Mohler & Mihalcea 2009). Sedangkan, kaedah yang lebih berkesan adalah diperlukan dalam usaha untuk menilai esei pendek

supaya penilaian yang lebih terperinci dan mendalam yang menggambarkan kualiti sebenar jawapan pelajar mampu diperolehi. Dalam mendapatkan penilaian teks pendek yang lebih optimum berbanding kaedah umum, kesetaraan teks merupakan antara penyelesaian yang lebih baik yang mana ianya melibatkan pengecaman dan perbandingan fitur-fitur antara dua teks (Mohler & Mihalcea 2009).

Terdapat kajian yang intensif dalam pengukuran kesetaraan ayat teks atau dokumen panjang, namun sebaliknya kajian pengukuran kesetaraan ayat atau teks pendek masih kurang (Islam & Inkpen 2008). Kaedah pengukuran kesetaraan teks pendek secara dasarnya boleh dikelaskan kepada tiga kategori utama; kaedah statistik, kaedah semantik dan kaedah hibrid. Kaedah statistik menggunakan pendekatan berpandukan-kekerapan kewujudan kata dan pendekatan berpandukan-korpus. Manakala kaedah semantik pula menggunakan pendekatan berpandukan-pengetahuan dan pendekatan berpandukan-vektor.

Di antara kesemua kaedah tersebut, kaedah kekerapan kewujudan kata adalah yang sering digunakan dalam menilai soalan teks-terbuka manakala kaedah lain lebih sesuai diaplikasikan ke atas soalan teks-tertutup. Soalan teks-terbuka memerlukan pelajar untuk menulis lebih panjang berkenaan dengan topik yang ditanya dengan hanya berdasarakan kepada pengalaman dan pengetahuan mereka (Diana Perez 2004). Sedangkan, soalan teks tertutup pula boleh dibahagikan kepada dua; soalan komprehensif dan soalan struktur. Bagi soalan struktur, jawapan yang dibina tidak boleh lebih dari lima ayat (Pulman & Sukkarieh 2005). Jadual 2.1 menunjukkan rumusan kaedah yang digunakan dalam PEBP.

Jadual 2.1 Kaedah-kaedah yang digunakan dalam PEBP.

Kajian	Kaedah	Teknik/Model	Pencapaian	Kelemahan
Shan et. al (2009)	Statistikal.	Model Kebarangkalian, Model Ruang Vektor, Kaedah Jarak Terubah dan Model N-gram.	Dokumen setara jika banyak kata yang sama.	Sangat efektif tetapi tidak sesuai untuk teks pendek.

bersambung…

…sambungan

Higgins & Burstein (2007), Burgess et. al (1998)	Kaedah statistik berpandukan-korpus.	*Latent Semantic Analysis* dan *Hypespace Analogues to Language*	Mengira kesetaraan menggunakan korpus bersaiz besar.	Saiz dimensi kata adalah terhad. Maklumat kekerapan-kewujudan kata menghasilkan ruang dimensi yang tinggi.
Ho et. al (2010)	Kaedah semantik berpandukan-pengetahuan.	Rangkaian Semantik *Wordnet*.	Mengambilkira makna terhampir daripada antara perbandingan dua ayat.	Implikasi susunan kata ke atas makna ayat.
Islam & Inkpen (2008)	Kaedah semantik berpandukan-vektor.	Pengesanan Plagiat Berpandukan-Kesetaraan (*SimPaD*).	Mengaplikasikan perbandingan ayat-ke-ayat.	Tidak mengambilkira susunan kata dalam ayat.
Sumathy (2016)	Kaedah hibrid.	Vektor semantik dan vektor susunan kata.	Kesetaraan ayat berpandukan kepada maklumat semantik dan maklumat susunan kata	Kedua-dua vektor diproses berasingan. Hasil kurang tepat.

2.2.1 Kaedah Statistik

Salah satu dari kajian paling awal dalam pengukuran kesetaraan ayat ialah pendekatan berpandukan-kekerapan kewujudan kata yang juga dikenali sebagai kaedah ‘beg kata’. Pendekatan ini sering digunakan dalam menilai soalan teks-terbuka manakala kaedah lain lebih sesuai diaplikasikan ke atas soalan teks-tertutup. Soalan teks-terbuka memerlukan pelajar untuk menulis lebih panjang berkenaan dengan topik yang ditanya dengan hanya berdasarkan kepada pengalaman dan pengetahuan mereka (Diana Perez 2004). Sedangkan, soalan teks tertutup pula boleh dibahagikan kepada dua; soalan komprehensif dan soalan struktur. Bagi soalan struktur, jawapan yang dibina tidak boleh lebih dari lima ayat (Pulman & Sukkarieh 2005).

Model Kebarangkalian, Model Ruang Vektor (MRV), Kaedah Jarak Terubah dan Model N-gram merupakan antara kaedah yang biasa digunakan dalam kekerapan kewujudan kata dan statistik kata (Shan et. al 2009). Kaedah ini adalah berdasarkan kepada anggapan bahawa sesuatu dokumen itu dikira sebagai setara jika dokumen-dokumen tersebut mempunyai banyak kata yang sama (Ning et. al 2011). MRV

berpandukan kepada Kekerapan Istilah-Kekerapan Dokumen Pembalikan (KIKDP). KIKDP, walaupun merupakan salah satu model yang sangat efektif dengan mengadaptasikan kaedah berasaskan statistik, masih tidak sesuai untuk teks pendek (Shan et. al 2009). Ini adalah kerana kekerapan kewujudan kata yang sama dalam dua teks pendek yang disetarakan adalah jarang dan mungkin juga tidak wujud, sedangkan kedua-dua teks tersebut adalah setara kerana penggunaan kata yang berbeza namun membawa makna yang sama secara semantiknya.

Kaedah statistik berpandukan-korpus biasanya mengira kesetaraan ayat berdasarkan maklumat statistik kata menggunakan korpus bersaiz besar. Kaedah ini seringkali digunakan dalam model *Latent Semantic Analysis* (LSA) (Higgins & Burstein 2007) dan *Hypespace Analogues to Language* (HAL) (Burgess et. al 1998). LSA melibatkan penggunaan Penghuraian Nilai Tunggal (PNT) ke atas matrik dokumen-secara-istilah dalam usaha untuk mengurangkan kedudukannya. Terdapat beberapa kelemahan LSA; [1] saiz dimensi kata berdasarkan matrik kandungan adalah terhad disebabkan kekangan yang wujud dalam had pengiraan PNT, seterusnya menjurus kepada berlakunya keadaan di mana beberapa kata ynag penting dari teks/ayat input mungkin tidak dimasukkan dalam ruang dimensi LSA dan ianya akan mengabaikan maklumat sintaksis daripada dua ayat (Li et. al 2006). Bagi HAL pula, maklumat kekerapan-kewujudan kata turut digunakan untuk menghasilkan ruang dimensi yang tinggi. Bagaimanapun, keputusan yang dihasilkan menunjukkan kemampuan HAL adalah tidak sebaik LSA dalam mengukur kesetaraan teks pendek. Kekurangan HAL berpunca daripada pembinaan matrik memori dan kaedah penjanaan vektor ayat. Matrik kata-ke-kata tidak menggambarkan maksud sebenar sesebuah ayat (Ning et. al 2011).

2.2.2 Kaedah Semantik

Sebagai perbandingan, kaedah semantik berpandukan-pengetahuan mengukur kesetaraan antara dua ayat berdasarkan kepada maklumat semantik yang diekstrak dari pangkalan data (Ho et. al 2010). Kebiasaannya, maklumat semantik ini diperolehi dengan mengambilkira makna terhampir dari perbandingan dua ayat tersebut. Ho et. al (2010) telah menghasilkan kaedah yang lebih optimum; perbandingan dua ayat

berdasarkan kepada makna sebenar dengan mengubahsuai pengukuran kesetaraan ayat berpandukan-korpus sedia ada kepada kaedah berpandukan-pengetahuan. Li et. al (2006) mengukur kesetaraan semantik antara ayat atau teks pendek berpandukan hubungan semantik dan maklumat susunan kata. Kesetaraan semantik diperolehi daripada pangkalan data pengetahuan dan korpus *Wordnet*. Seterusnya, mereka juga mempertimbangkan implikasi susunan kata ke atas makna ayat.

Kaedah semantik berpandukan-vektor biasanya digunakan dalam sistem DSM, di mana dokumen yang paling relevan dengan teks input ditentukan dengan mewakilkan sebuah dokumen sebagai vektor kata dan teks input dipadankan dengan dokumen yang setara dalam pangkalan data dokumen melalui matrik kesetaraan (Islam & Inkpen 2008). Antara kajian lanjutan bagi kaedah berpandukan-vektor ini adalah menggunakan Pengindeksan Rawak untuk mencari dokumen dalam ruang semantik (Higgins & Burstein 2007). Pengindeksan Rawak menghasilkan vektor semantik untuk setiap kata dalam korpus, di mana ianya seterusnya dibandingkan dengan vektor bagi kata yang lain menggunakan piawaian metrik kesetaraan kosinus (Lee 2010). Selain itu, terdapat juga kajian lanjutan lain dalam kaedah berpandukan-vektor yang menjurus kepada Pengesanan Plagiat Berpandukan-Kesetaraan (*SimPaD*) yang mengaplikasikan perbandingan ayat-ke-ayat (Gustafson et. al 2008). Kaedah ini bersandarkan kepada faktor kolerasi kata pra-kira untuk mengenalpasti kesetaraan ayat-ke-ayat dan akhirnya kadar persamaan bagi mana-mana dua dokumen dikesan sebagai salinan plagiat. Walaubagaimanapun, memandangkan SimPaD tidak mengambilkira aturan kata dalam ayat, pengukuran kesetaraan teks pendek mungkin kurang jitu.

Beberapa kajian terbaru lain turut menunjukkan penggunaan kaedah pengukuran ayat berpandukan-semantik mampu mencapai keputusan yang baik. Wang et. al (2016) mencadangkan sebuah model yang mengambilkira kedua-dua kesetaraan dan ketidaksetaraan dengan menghurai dan membentuk semantik leksikal ke atas ayat. Model tersebut mewakilkan setiap kata sebagai vektor dan mengira vektor padanan semantik untuk setiap kata berpandukan kepada semua kata dalam ayat yang lain. Seterusnya setiap vektor kata dihuraikan kepada komponen yang setara dan tidak setara berdasarkan kepada vektor padanan semantik. Setelah itu, model C*onvolutional*

Neural Network dua-saluran digunakan untuk mendapatkan fitur-fitur dengan menghurai komponen yang setara dan tidak setara. Akhir sekali, pengukuran kesetaraan dikira berdasarkan vektor fitur terhurai. Hasil eksperimen menunjukkan pencapaian yang dicapai adalah setara dengan pencapaian terkini dalam tugas pemilihan ayat jawapan dan pencapaian yang agak baik bagi tugas penentuan parafrasa.

Li & Li (2015) membuktikan dengan menggunakan algoritma yang berpandukan kepada struktur sintaksis dan melakukan pengukuran hubungan semantik ke atas struktur sintaksis tersebut berjaya meningkatkan tahap keberkesanan pengukuran kesetaraan ayat. Lebih menarik lagi, hanya dengan mengubahsuai dan menggabungkan beberapa kaedah pengukuran hubungan semantik, ianya berjaya meningkatkan hasil ketepatan kesetaraan (Ptáček 2012).

2.2.3 Kaedah Hibrid

Manakala itu, kaedah hibrid pula menggabungkan pendekatan berpandukan semantik, korpus, ontologi dan hubungan (Sumathy & Chidambaram 2016). Li et. al (2006) menunjukkan satu kaedah pengukuran kesetaraan ayat berpandukan kepada maklumat semantik dan maklumat susunan kata yang wujud dalam sesebuah ayat. Mulanya, kesetaraan semantik diperolehi daripada perbandingan antara vektor semantik mentah dan vektor semantik dengan pangkalan data leksikal dan korpus. Seterusnya, kesetaraan susunan kata akan dihasilkan melalui perbandingan kedua-dua set vektor tersebut. Akhirnya, kesetaraan ayat diukur dengan menggabungkan kesetaraan semantik dan kesetaraan susunan kata.

Kaedah hibrid merupakan antara kaedah yang seringkali digunakan dalam menilai kesetaraan ayat esei pendek (Sumathy & Chidambaram 2016, Li et. al 2006, Pawar & Mago 2018). Malah kajian terbaru dalam mengukur kesetaraan ayat masih mengaplikasikan kaedah hibrid ini. Pawar & Mago (2018) mengira kesetaraan antara kata berpandukan pendekatan berpandukan-sisi. Kandungan maklumat daripada pangkalan data leksikal *Wordnet* tersebut dipercayai mampu mempengaruhi pengukuran kesetaraan dalam domain spesifik. Vektor semantik yang mengandungi

kesetaraan antara kata akan membentuk ayat dan digunakan untuk pengiraan kesetaraan ayat. Vektor susunan kata juga akan dibentuk untuk mengira impak struktur sintaksis ke atas ayat. Kesetaraan ayat dikira berdasarkan kedua-dua vektor semantik dan vektor susunan kata tersebut.

2.3 ANALISA TERPERINCI KAEDAH PEPB SEDIA ADA

Dalam kajian yang melibatkan pengukuran kesetaraan semantik bagi terjemahan frasa dan ayat pendek dari Bahasa Arab ke Bahasa Inggeris, Teknik Terjemahan Mesin dan Kamus telah digunakan (Salha Alzahrani 2016). Algoritma Purata Terjemahan-Maksimum akan menggunakan set frasa yang dihasilkan oleh Teknik Berpandukan-Kamus. Vektor frasa dan kata nama-kata kerja yang diperoleh daripada Teknik Terjemahan Mesin juga pula digunakan untuk mengira kesetaraan semantik. Walaubagaimanapun, disebabkan pengukuran kesetaraan ayat hanya melibatkan pengukuran kesetaraan semantik tanpa melihat kepada struktur ayat, maka wujud isu kekaburan kata.

Kadupitiya et. al (2016) menggunakan Teknik Pengukuran Kesetaraan Semantik (pengukuran kesetaraan berpandukan-korpus dan berpandukan-pengetahuan). Teknik tersebut menggunakan konsep kesetaraan hubungan semantik sepertimana taksonomi *Wordnet* selain dari menggunakan maklumat susunan perkataan. Namun begitu, kajian menunjukkan hasil yang diperolehi boleh dioptimumkan lagi jika masalah kekaburan kata dapat diatasi dengan mempertimbangkan kata sekeliling bagi mendapatkan maklumat konteks ayat.

Kajian terbaru ke atas bahasa lain turut melibatkan penentuan fitur statistik dan semantik dalam pengukuran kesetaraan ayat dalam Bahasa Portugis. Empat fitur yang agak asas digunakan iaitu KIKDP, *Word2Vector*, Kaedah Matrik Binari dan panjang ayat (Anderson Pinheiro et. al 2017). KIKDP adalah kaedah statistik untuk mengukur tahap signifikan sesebuah kata yang wujud dalam ayat (Salton & Yang 1973). *Word2Vector* merupakan model *unsupervised* untuk menjana perwakilan vektor bagi setiap kata dalam set kata yang bertujuan untuk mengukur kesetaraan semantik antara kata (Rumelhart et. al 1988). Kaedah Matrik Binari pula menggunakan Kaedah

Berpandukan-Matrik untuk mengira kesetaraan antara ayat yang ditentukan berdasarkan kesetaraan antara kata (Ferreira et. al 2016). Fitur terakhir yang juga digunakan oleh Zhao et. al (2014) dan Bjerva et. al (2014) ialah panjang ayat. Ianya diukur berdasarkan bilangan kata dalam ayat yang paling pendek dibahagikan dengan bilangan kata dalam ayat yang paling panjang. Bagi kaedah ini, kata henti dibuang terlebih dahulu. Walaubagaimanapun, penggunaan kaedah statistik dan fitur asas seperti panjang ayat tidak mampu mengukur kesetaraan ayat dari segi konteks.

Wang et. al (2016), dengan kajian terbarunya mengukur kesetaraan ayat Bahasa Thai berpandukan kaedah yang seringkali digunakan pada Bahasa Inggeris iaitu Struktur Sintaksis dan Vektor Semantik. Pengukuran kesetaraan ayat menggunakan penanda golongan kata dan kebergantungan penanda tersebut untuk mengira kesetaraan struktur sintaks, seterusnya mengukur kesetaraan semantik ayat menggunakan *Word2Vector*. Kajian juga menunjukkan hasil yang diperolehi mendapat kesetaraan yang lebih tepat kerana kaedah pengukurannya tidak hanya menilai dari segi semantik ayat, tetapi juga maklumat struktur ayat. Namun begitu, kajian lanjutan masih perlu dibuat dengan mempertimbangkan struktur ayat menggunakan kaedah linguistik yang dijangka mampu mengekstrak peranan yang dimainkan oleh argumen dalam ayat.

Walaubagaimanapun, pengukuran kesetaraan ayat berdasarkan strukur ayat atau susunan kata masih belum mampu menyelesaikan masalah kekaburan kata. Masalah kekaburan kata boleh diatasi jika peranan kata sekeliling diambilkira untuk mendapatkan sebahagian maklumat konteks (Li et. al 2006).

2.4 ANALISA KAEDAH KAJIAN TERDAHULU

Kaedah statistik tidak selalunya mampu mengenalpasti padanan yang sempurna tanpa hubungan atau konsep yang jelas antara dua ayat tabii. Beberapa pendekatan digunakan untuk menangani masalah ini dengan menentukan susunan kata dan penilaian vektor semantik, namun pendekatan tersebut sukar untuk membandingkan ayat yang mempunyai struktur sintaksis yang kompleks yang dibina berdasarkan

penggunaan kata dan ayat yang panjang menggunakan corak nahu yang pelbagai (Lee et. al 2014).

Bagi menyelesaikan masalah tersebut, para pengkaji menggunakan kaedah semantik (Lee et. al 2012, Mandreoli et. al 2002). Kaedah ini menggunakan rangkaian semantik seperti *Wordnet*, Model Ruang Vektor dan Korpus Statistik untuk mengira kesetaraan semantik antara kata menggunakan kaedah pengukuran yang berbeza. Walaubagaimanapun, kaedah semantik ini mengukur kesetaraan ayat hanya berpandukan kepada kesetaraan semantik antara kata, di mana maklumat sintaksis dan pengetahuan semantik yang lain seperti kelas semantik dan peranan tematik, diabaikan (Wafa Wali et. al 2017).

Bagi mengatasi isu tersebut, para pengkaji mencadangkan kaedah hibrid untuk mengira kesetaraan ayat dengan mempertimbangkan kedua-dua maklumat semantik dan sintaksis. Namun begitu, kaedah hibrid ini mungkin terdapat beberapa kekurangan iaitu pengukuran semantik dibuat secara berasingan di mana kesetaraan semantik dikira berdasarkan kepada kesetaraan semantik kata, manakala padanan frasa, susunan kata dan kekerapan kewujudan kata dikira untuk kesetaraan sintaksis. Malah, beberapa fitur pengetahuan tidak dipertimbangkan dalam pengukuran kesetaraan ayat seperti peranan tematik, kelas semantik dan hubungan antara aras sintaksis dan semantik berdasarkan predikat semantik (Wafa Wali et. al 2017). Apabila dua ayat yang mempunyai struktur sintaksis yang sama (subjek + kata kerja + objek) tetapi kelas semantik kata kerja atau tersebut berbeza, pasangan ayat tersebut adalah setara secara sintaksisnya berdasarkan kaedah hibrid, sedangkan sebenarnya kedua-dua ayat tersebut adalah berbeza menurut pakar manusia. Sebagai contoh, ayat ‘Ali membaca buku’ dan ayat ‘Ali mempunyai buku’, mempunyai struktur sintaksisnya yang sama dan hubungan semantik bagi setiap argumen (subjek (kata nama) + kata kerja + objek (kata nama)) yang wujud dalam kedua-dua ayat tersebut, namun berbeza dari segi peranan tematik yang dimainkan oleh kata kerja, seterusnya menyebabkan kedua-dua ayat tersebut sebenarnya tidak setara langsung.

Lee et. al (2014) telah menggunakan ontologi berpandukan-korpus untuk mengira hubungan kesetaraan antara dua kata dan petua nahu dalam usaha untuk

mengenalpasti konteks ayat. Wafa Wali et. al (2017) pula menggunakan teknik pengetahuan semantik bagi menentukan kesetaraan semantik dan teknik pengetahuan sintaks-semantik bagi mengurangkan masalah kekaburan kata. Kedua-dua kaedah tersebut merupakan kaedah linguistik yang mengaplikasikan petua nahu dalam menangani isu kekaburan kata dengan cara menentukan konteks ayat.

2.5 KAEDAH UMUM KESETARAAN AYAT

Beberapa kajian terkini masih menggunakan kaedah pengukuran kesetaraan ayat untuk menilai esei pendek bersepadu (Twinkle et al 2015; Rei & Cummins 2016; Alla Alrehily et. al 2018). Kesetaraan ayat merupakan kaedah yang seringkali dikaji dalam Pemprosesan Bahasa Tabii kerana ianya banyak digunakan dalam sistem soal-jawab, perlombongan teks, peringkasan teks, pengesanan plagiat, penilaian jawapan pelajar dalam bidang teknologi pendidikan atau penilaian kualiti terjemahan bagi sistem terjemahan automasi (Dan Ștefănescu et. al 2014). Menurut Wang & Brown (2007), kesetaraan ayat dikira berdasarkan kesetaraan kata. Namun, pendapat ini disangkal oleh Dan Ștefănescu et. al (2014) iaitu menurutnya kesetaraan ayat bukan sekadar perbandingan kata, namun ianya melibatkan kajian ke atas penggunaan pelbagai kaedah lain supaya kesetaraan ayat mencapai kesetaraan peringkat semantik. Malah beberapa kajian lain menggunakan konsep kesetaraan semantik dan kesetaraan ayat untuk mencari kadar padanan antara skema jawapan dan jawapan pelajar yang lebih tepat (Alla Alrehily et. al 2018).

Lee et. al (2014) telah membentangkan kajiannya yang mengandungi algoritma kesetaraan berpandukan korpus dan nahu untuk ayat bahasa tabii. Bagi mencapai hasil kesetaraan yang lebih baik, algoritma kesetaraan ayat tersebut menggunakan fitur-fitur tertentu yang wujud dalam ontologi berpandukan-korpus dan petua nahu. Pawar & Mago (2018) pula membentangkan kajiannya berdasarkan pengiraan kesetaraan semantik antara dua kata, perenggan atau ayat. Sebagai perbandingan, algoritma tersebut dimulakan dengan menyah-kabur kesemua ayat untuk memastikan makna yang tepat bagi setiap kata. Dengan menggunakan pangkalan data leksikal, pendekatan berpandukan-pinggiran (*edge*) diperkenalkan untuk mengira kesetaraan semantik antara kata dan ayat (Li et. al 2006). Jadual 2.2

menunjukkan beberapa kaedah kesetaraan yang telah digunakan dalam penilaian esei pendek bersepadu sehingga kini yang membuktikan kaedah ini masih relevan untuk membuat penilaian esei pendek menghampiri penilaian manusia.

Jadual 2.2 Kaedah umum pengukuran kesetaraan ayat bagi esei pendek.

Kajian	Kaedah	Pencapaian
Higgins & Burstein (2007)	Pengindeksan Rawak, kesetaraan semantik berpandukan-vektor.	Pengukuran kesetaraan ayat dibuat dengan menjumlahkan Pengindeksan Rawak dan *Latent Semantic Analysis*. Set data yang digunakan adalah daripada akhar *San Mercury News*. Ketepatan maksimum yang diperolehi adalah 72.1%.
Khaled Abdalgader & Andrew Skabar (2010)	Penyahtaksaan Makna Kata dan Pengembangan Sinonim.	Teknik Penyahtaksaan Makna Kata dan Pengembangan Sinonim digunakan untuk menyediakan konteks semantik yang lebih optimum untuk mengukur kesetaraan ayat. Min kolerasi manusia adalah 82.5% bagi set data 30 ayat.
Ali Muftah, Mohd Juzaiddin (2013)	Kata Biasa, Subjujukan Biasa Terpanjang dan Jarak Semantik.	Mengenalpasti kesetaraan ayat antara jawapan pelajar dan model jawapan dijana menggunakan gabungan Kata Biasa, Subjujukan Biasa Terpanjang dan Jarak Semantik. Set data daripada 40 soalan yang dijawab oleh tiga orang pelajar. Menghasilkan 82% kolerasi dengan penilaian manusia.
Lee et. al (2014)	Algoritma Kesetaraan Berpadukan-Nahu dan Korpus Semantik.	Menggunakan Ontologi Berpandukan-Korpus dan Petua Nahu utuk mengukur kesetaraan semantik antara ayat yang mengintegrasikan penilaian kata-ke-kata. Set data daripada *Microsoft Research Paraphrase Corpus*. Ujian ketepatan dan dapatan masing-masing setinggi 0.91 dan 0.98 bagi nilai ambang 0.6-1.0.
Dan Ştefănescu et. al (2014)	Kesetaraan Pecahan dan Maklumat Fitur.	Menggunakan konstituen sintaksis untuk mengukur kesetaraan ayat berdasarkan kepada Kesetaraan Pecahan dan Maklumat Fitur. Set data daripada *Microsoft Research Paraphrase Corpus*. Ujian ketepatan adalah 0.742 dan ukuran-f adalah 0.821.

bersambung…

…sambungan

Pearce (2015)	Algoritma SARUMAN.	Menggunakan model matematik yang menggabungkan komponen lingustik dinamakan SARUMAN untuk mengenalpasti kesetaraan ayat. Set data dijana daripada set data *Microsoft Research Paraphrase Corpus* and Kesetaraan Sematik (STASIS). Hasil ujian Pearson ialah 0.906 dan Spearman ialah 0.802.
Feddy Pribadi et. al (2017)	Pekali Jaccard, Pekali Dadu dan Pekali Kosinus.	Menggunakan Kaedah Pekali Jascard, Pekali Dadu dan Pekali Kosinus untuk mengira kesetaraan kata yang sama dalam ayat. Set data daripada 2160 soalan jawapan-pendek. Kolerasi antara penilaian mesin dan penilaian manusia ialah 75.9%.
El Moatez Billah et. al (2018)	Model Berpandukan-Penterjemahan Mesin dan Kesetaraan Semantik Bahasa Tunggal.	Menggunakan Model Berpandukan-Penterjemahan Mesin diikuti dengan Kesetaraan Semantik Bahasa Tunggal berdasarkan kepada kata terbenam. Set data yang digunakan adalah daripada *International Workshop on Semantic Evaluation (SemEval-2017).* Hasil kolerasi iaitu 77.39.
Alla Alrehily et. al (2018)	Mesin Vektor Sokongan	Menggunakan konsep kesetaraan semantik dan kesetaraan ayat untuk mencari padanan antara jawapan pakar dan jawapan pelajar untuk setiap soalan. Set data adalah daripada 100 soalan merangkumi 5 soalan objektif dan 50 soalan subjektif. Keputusan kejituan, dapatan dan ukran-f masing-masing 0.9072, 0.8900 dan 0.9412.
Pawar & Mago (2018)	Pendekatan Berpandukan-Pinggiran dan Rangkaian Semantik	Kesetaraan antara kata dikira berdasarkan kepada pendekatan Berpandukan-Sempadan yang telah dibangunkan. Maklumat Fitur daripada korpus digunakan untuk mengukur kesetaraan domain. Vektor semantik yang mengandungi kesetaraan antara kata untuk pengukuran kesetaraan ayat.

bersambung…

…sambungan

		Set data diperoleh daripada *Pilot Short Text Semantic Similarity Benchmark Data Set* oleh Shea, J. O. et al (2008). Hasilnya, pekali kolerasi Pearson adlaah 0.8753 dan untuk kesetaraan ayat, kolerasi yang diperolehi adalah 0.8794.
Uswatun Hasanah et. al (2019)	Subjujukan Biasa Terpanjang, Pekali Jaccard, Pekali Dadu dan Pekali Kosinus	Menggunakan kaedah padanan berpandukan-jujukan dan kaedah padanan berpandukan-kata untuk mengukur kesetaraan ayat. Set data bahasa Indonesia yang terdiri daripada 7 soalan dengan 34 maklumbalas pakar sebagai skema jawapan dan sebanyak 256 jawapan pelajar yang setiap satunya dijawab oleh 32 orang pelajar. Penggunaan kaedah LCS menghasilkan kolerasi 66.38%.
Farah Nadeem et. al (2019)	Model Neural Berpandukan-Hujahan	Menggunakan model neural dengan kebergantungan ayat-silang dan latihan berpandukan-hujahan untuk mengkira kesetaraan ayat. Set data daripada 12,100 esei TOEFL. Kolerasi maksimum adalah 72.9%.

Jadual 2.2 menunjukkan beberapa kajian yang telah dibuat menggunakan kaedah pengukuran kesetaraan ayat bagi bagi menilai esei pendek. Walaupun ramai pengkaji mengaplikasikan kaedah statistikal, namun masih ramai lebih menyakini kaedah linguistik dalam mengukur kesetaraan ayat tersebut (Lee et. al 2014, Dan Ştefănescu et. al 2014, Pawar & Mago 2018, El Moataz Billah et. al 2018). Pendekatan dapatan maklumat tradisional seperti model vektor, LSA, HAL atau pendekatan berpandukan-ontologi, yang menekankan perbandingan kesetaraan konsep berbanding kekerapan kewujudan kata/frasa, mungkin tidak semestinya dapat mengenalpasti padanan yang tepat disebabkan tiadanya hubungan atau konsep berkaitan yang jelas antara dua ayat bahasa tabii (Lee et. al 2014).

2.6 PENGUKURAN KESETARAAN AYAT BAHASA MELAYU

Sementara pelbagai usaha dilakukan untuk mengukur kesetaraan ayat untuk Bahasa Inggeris, hanya sebahagian penyelidik memfokuskan terhadap Bahasa Melayu. Antara faktor utama adalah kerana kekurangan alatan asas dalam memproses Bahasa Melayu melalui kaedah linguistik (penanda golongan kata, semantik, ontologi). Shahrul Azman et. al (2007) secara spesifiknya cuba mengukur kesetaraan ayat Bahasa Melayu berdasarkan pendekatan berpandukan-vektor. Ianya menggunakan Kamus Bahasa Melayu Pra-Proses dan Kaedah Berpandukan-Pengiraan Pinggir Bertindih untuk mengira kesetaraan semantik kata-ke-kata, pada peringkat awal. Hasil daripada proses tersebut akan digunakan untuk mengenalpasti kesetaraan ayat semantik menggunakan pendekatan untuk Bahasa Inggeris yang diubahsuai untuk membangunkan vektor semantik dan vektor susunan perkataan. Akhirnya, hasil tambah kedua-dua vektor adalah merupakan hasil pada pengukuran kesetaraan ayat. Daripada kajian yang dibuat, hasil akhir yang diperolehi adalah menggalakkan dan konsisten jika penggunaan pendekatan ini dibandingkan dengan penilaian manusia. Sungguhpun begitu, terdapat beberapa isu yang wujud. Berdasarkan eksperimen di peringkat awal, dua ayat yang sememangnya setara telah dinilai oleh sistem hanya sebanyak 57.9% setara setelah mengabaikan elemen morfologi. Namun sebaliknya, tahap kesetaraan meningkat kepada 90.2% jika elemen tersebut diambilkira. Oleh itu, pengukuran kesetaraan ayat menggunakan kaedah kata-ke-kata seharusnya perlu ditambahbaik dengan menggunakan kaedah-kaedah lain seperti kaedah berpandukan korpus kekerapan-kewujudan-kata dan rangkaian semantik yang mampu mengukur kesetaraan ayat berdasarkan konteks sebenar ayat tersebut.

Kajian lanjutan dalam mengukur kesetaraan ayat Bahasa Melayu adalah berpandukan kepada kaedah baru iaitu Teknik Pola *Grammar*. Teknik tersebut mengukur kesetaraan menggunakan perbandingan hubungan nahu ayat (Mohd Juzaiddin et. al 2008). Pada peringkat permulaan, ianya akan mengekstrak hubungan nahu dalam ayat dan kemudiannya akan memadankannya dengan ayat-ayat yang lain. Kesetaraan ditentukan berdasarkan pada kesemua subjek, kata kerja dan objek adalah setara. Jika didapati hanya subjek yang setara, ianya seterusnya akan merujuk pada kata kerja yang sinonim dalam Tesaurus Bahasa Melayu. Hasil yang diperolehi

menunjukkan purata perbezaan penilaian antara kaedah ini berbanding dengan penilaian manusia adalah serendah 0.049, 0.028 dan 0.12 diuji ke atas tiga set soalan. Dengan aras signifikan 0.05, ianya boleh disimpulkan bahawa penilaian yang dibuat oleh teknik ini adalah menghampiri penilaian manusia. Malah, Teknik Pola *Grammar* ini boleh dioptimumkan dengan mengintegrasikan semantik leksikal berpandukan petua hujahan ke atas kata fungsi dalam usaha untuk mendapatkan saiz domain berkaitan.

Usaha ini diteruskan oleh Suhaimi et. al (2011) menggunakan pangkalan data frasa berpandukan-petua dan pangkalan pengetahuan konteks sinonim untuk mendapatkan perkataan khusus yang sinonim dalam ayat Bahasa Melayu. Penanda golongan kata Bahasa Melayu digunakan untuk menanda jenis leksikal ke atas setiap perkataan yang ditokenkan dalam teks yang dimasukkan. Kemudian, pengelas perkataan berpandukan-pergantungan akan mengenalpasti padanan corak antara perkataan yang telah ditanda dengan petua dalam pangkalan data frasa berpandukan-petua. Bagi menentukan konteks perkataan dalam teks Bahasa Melayu, modul penentuan kesetaraan berpandukan-konteks dilaksanakan. Dengan mengaplikasikan petua inti dan penerang, ianya akan menentukan konteks perkataan tersebut dengan membuat carian dalam pangkalan pengetahuan konteks sinonim. Bagaimanapun, pendekatan yang lebih komprehesif dalam mengenalpasti konteks perkataan dalam setiap frasa menggunakan petua inti dan penerang adalah sangat diperlukan.

2.7 JUSTIFIKASI PEMILIHAN TEKNIK

Berdasarkan kajian terdahulu yang dibuat, dua isu utama dalam mengukur kesetaraan ayat dengan lebih berkesan ialah dengan mengenalpasti konteks ayat dan pemilihan kaedah pengukuran hubungan semantik yang mampu mengukur kesetaraan semantik dengan lebih tepat. Ayat yang diukur kesetaraannya pula tidak seharusnya diproses berasingan, sebaliknya diproses secara bersilang antara kedua-dua teknik tersebut.

Konteks ayat tidak mampu diukur berdasarkan susunan kata atau struktur sintaksis. Oleh itu, kajian ini menggunakan pendekatan linguistik, iaitu Petua Peranan Tematik sebagai usaha untuk menentukan peranan tematik yang dimainkan oleh

argumen dalam ayat seterusnya konteks ayat dikenalpasti. Kelas semantik dan peranan tematik untuk setiap argumen dalam sesebuah ayat mampu menyediakan maklumat berkenaan hubungan antara kata dan peranan yang dimainkan dalam menentukan makna sesebuah ayat (Wafa Wali et. al 2017).

Bagi pengukuran kesetaraan *synset* (kata/argumen) dalam ayat pula, Teknik Rangkaian Semantik berpandukan-pengetahuan berdasarkan senibina *Wordnet* dipilih menggunakan kaedah pengukuran kesetaraan yang lebih baik. Terdapat tiga jenis pendekatan kesetaraan semantik yang berbeza iaitu pendekatan kesetaraan kosinus, pendekatan berpandukan laluan (Kaedah *Wu & Palmer* (wup) dan Kaedah Laluan Terpendek) dan pendekatan berpandukan fitur. Kajian menunjukkan pendekatan berpandukan fitur menghasilkan kesetaraan semantik antara ayat lebih baik (Sravanti & Srinivasu 2017). Namun begitu, pendekatan tersebut melibatkan pengiraan kesetaraan yang agak kompleks.

Pengukuran kesetaraan semantik terkini kebiasannya mengaplikasikan ontologi atau perwakilan taksonomi bagi konteks. Pengukuran kesetaraan semantik wup merupakan salah satu daripada pengukuran tersebut yang dikategorikan sebagai ringkas dan berprestasi tinggi, namun ia boleh menghasilkan keputusan yang kurang tepat disebabkan oleh dua konsep dalam hierarki ontologi yang sama mungkin menunjukkan tahap kesetaraan yang lebih rendah daripada dua konsep yang berada dalam hierarki yang berbeza. Namun begitu, Guessoum et. al (2016) telah mengubahsuai pengukuran kesetaraan semantik wup sedia ada mengekalkan fitur-fitur perlaksanaannya yang ringkas dan berprestasi tinggi namun meningkatkan tahap ketepatan pengukuran kesetaraannya menghampiri penilaian manusia. Pengubahsuaian tersebut menepati empat kriteria penting pengukuran kesetaraan iaitu bukan-negatif ($Sim(A,B) \geq 0$), identiti ($Sim(A,A)=Sim(B,B)=1$), simetri ($Sim(A,B)=Sim(B,A)$) dan unik ($Sim(A,B)=1 \rightarrow A=B$). Kelebihannya ialah bahawa semua konsep dalam hierarki yang sama mestilah lebih setara antara satu sama lain berbanding konsep lain dalam hierarki yang berbeza dan persamaan antara konsep dalam hierarki yang sama juga bergantung pada jarak antara konsep-konsep ini.

2.8 PETUA SUBJEK-PREDIKAT

Petua Subjek-Predikat merupakan petua umum dalam pembangunan ayat. Ayat majmuk dibina daripada dua klausa bebas (satu subjek dan satu predikat) yang mana kedua-duanya sama penting dan boleh hadir dengan sendirinya tanpa kehadiran klausa yang satu lagi (Lingard 2017). Subjek merupakan konstituen dalam sesuatu ayat yang terdiri daripada satu perkataan atau beberapa perkataan yang berfungsi sebagai frasa nama dan menjadi unsur yang diterangkan. Predikat pula merupakan kumpulan perkataan yang tergolong dalam satu frasa dan menjadi unsur yang menerangkan subjek.

Kebanyakan kajian yang mengaplikasikan kaedah susunan kata dalam membuat penilaian esei pendek bersepadu mempraktikkan Petua Subjek-Predikat (Li et. al 2006, Lee et. al 2012). Bagi Sukkarieh & Blackmore (2009), apabila sesebuah ayat mempunyai subjek tetapi tidak wujud sebarang objek, maka subjek tersebut dianggap sebagai objek untuk tujuan padanan. Rencam nahu dinilai berdasarkan Petua Subjek-Predikat dan tahap komplesiti ayat dikira setara dalam menentukan tahap pembacaan ayat (Shermis et. al 2010). Komputer menilai sesebuah esei menggunakan model statistikal dan linguistik bergantung kepada elemen-elemen yang ditentukan oleh penilai manusia berdasarkan kepada sampel penulisan (skema jawapan) (Shermis 2010). Yamamoto et. al (2018) dalam kajiannya menjadikan kebergantungan kepada subjek dan predikat sebagai sebahagian daripada 25 item dalam algoritma penilaiannya.

Walaubagaimanapun, setiap bahasa mempunyai petua nahunya tersendiri dalam pembinaan ayat yang gramatis. Chang et. al (2019) dalam kajiannya melibatkan penilaian esei Bahasa Cina bersepadu menyatakan bahawa struktur ayat Bahasa Cina dan Inggeris terdiri daripada tiga elemen asas tetapi terdapat variasi antara dua bahasa tersebut dari segi ragam ayat. Sebagai contoh, kata kerja dan kata adjektif dalam Bahasa Cina boleh bertindak sebagai subjek, tetapi dalam Bahasa Inggeris, kata kerja mesti dalam bentuk kata kerja benda atau kata kerja waktu untuk bertindak sebagai subjek. Kekangan petua ini juga diaplikasikan ke atas predikat dan objek. Malah, berdasarkan petua transformasi struktur ayat dalam Bahasa Cina, posisi objek boleh

dialihkan kepada sebelum objek, namun ianya tidak mungkin bagi struktur ayat Bahasa Inggeris.

Kajian ini menekankan penggunaan kaedah linguistik dalam membuat penilaian esei pendek. Penggunaan Petua Peranan Tematik, dalam mengenalpasti peranan yang dimainkan oleh subjek dan objek dengan kehadiran kata kerja tertentu dalam ayat, memberi hasil pengukuran kesetaraan antara dua ayat berpandukan konteks ayat berbanding pengukuran argumen subjek dan objek berdasarkan posisi keduanya.

2.9 TEKNIK POLA *GRAMMAR*

Istilah pola merujuk kepada ‘ragam’ iaitu ‘ragam ayat’ (Mohd Juzaiddin et. al 2006). Pola *Grammar* merupakan teknik untuk mengekstrak fitur sintaksis dan hubungan nahu daripada struktur ayat bahasa melayu (Mohd Juzaiddin 2008). Pembentukkan sesebuah ayat Bahasa Melayu ditentukan oleh posisi argumen subjek dan predikat. Berdasarkan Asmah (2009), terdapat tujuh pola nahu Bahasa Melayu yang digariskan (Mohd Juzaiddin et. al 2006):

i. Pelaku + Perbuatan

ii. Pelaku + Perbuatan + Pelengkap

iii. Perbuatan + Pelengkap

iv. Diterangkan + Menerangkan

v. Digolong + Penggolong

vi. Pelengkap + Perbuatan + Pelaku

vii. Pelengkap + Perbuatan

Kajian ini menjadikan hasil kesetaraan ayat dalam penilaian esei pendek menggunakan Teknik Pola *Grammar* sebagai perbandingan dan garis dasar disebabkan dua faktor utama iaitu:

i. Teknik Pola *Grammar* adalah sebuah teknik penilaian esei pendek berbentuk linguistik. Teknik ini mengenalpasti posisi argumen subjek, kata kerja dan objek dalam ayat ringkas, ayat majmuk dan ayat kompleks. Seterusnya, setiap argumen tersebut dipadankan dengan argumen yang wujud dalam skema jawapan. Pendekatan yang hampir sama digunakan menggunakan Petua Peranan Tematik. Namun begitu, sebagai perbezaanya, petua ini membuat perbandingan dari segi konteks argumen dalam ayat.

ii. Petua Peranan Tematik dalam kajian ini menggunakan set data yang sama digunakan oleh Mohd Juzaiddin (2008) yang mengaplikasikan Teknik Pola *Grammar*. Bagi menghasilkan perbandingan hasil yang boleh dipercayai, satu set data latihan dan tiga set data ujian yang sama telah digunakan dan perbandingan hasil yang adil telah dihasilkan.

2.10 PERANAN TEMATIK

Peranan Tematik berfungsi untuk mengenalpasti peranan yang dimainkan oleh kata kerja dalam penandaan tematik Bahasa Melayu (Ramli 2006).

2.10.1 Pemilihan Kategori Sintaksis

Pemilihan kategori sintaksis merujuk kepada kaedah untuk mengenalpasti kategori yang dipilih bagi sesuatu kata kerja (KK). Ianya menentukan bilangan frasa nama (FN) yang perlu wujud dalam frasa kerja (FK) tersebut. Kata kerja boleh dibahagikan kepada tiga kategori iaitu kata kerja transitif (KKT), kata kerja dwitransitif (KKD) dan kata kerja tak transitif (KKTT). Bagi frasa kerja yang mengandungi kata kerja transitif seperti hurai, ianya memerlukan satu objek langsung frasa nama sebagai penyambut. Oleh itu, kata kerja tersebut memerlukan satu komplemen. Bagi frasa kerja yang mempunyai kata kerja dwitransitif yang juga disebut sebagai kata kerja transitif ganda seperti ‘tukar’, ianya memerlukan dua penyambut iaitu dua objek langsung frasa nama atau satu objek langsung frasa nama dan satu objek tak langsung frasa pelengkap (FP). Maka, dua komplemen diperlukan. Sebaliknya, bagi frasa kerja yang mengandungi kata kerja tak transitif seperti ‘jalan’, ianya tidak memerlukan sebarang komplemen.

Bagi kata kerja khas seperti kata pemeri ‘adalah’ dan ‘ialah’ serta ‘merupakan’, satu objek langsung frasa nama diperlukan sebagai penyambut.

i. Penghurai menghurai token

ii. Nahu bebas konteks menukar token kepada frasa nahu

iii. Pemproses sedang berjalan

iv. Proses tersebut adalah penganalisis leksikal

v. Pengkompil ialah penterjemah

vi. Penganalisis leksikal merupakan komponen pengkompil

Ayat (i) hingga (vi) merupakan contoh-contoh ayat yang menggunakan samada kata kerja transitif, kerja dwitransitif dan kata kerja tak transitif. Kategori tersebut ditentukan berpandukan kepada bentuk imbuhan dalam Jadual 2.3.

Jadual 2.3 Bentuk Imbuhan Jenis Kata Kerja

Jenis Kata Kerja	**Bentuk Imbuhan**
Kata Kerja Transitif	Imbuhan meN…kan
	Imbuhan meN…i
	Imbuhan memper…kan
	Imbuhan memper…i
	Imbuhan ber…kan
	Imbuhan meN… (diikuti dengan kata nama)
Kata Kerja Dwitransitif	Imbuhan meN… (diikuti dengan kata kepada)
Kata Kerja Tak Transitif	Imbuhan ber…
	Imbuhan ter…
	Imbuhan di…
	Imbuhan di…kan/i
	Imbuhan diper…kan/i
	Imbuhan meN… (diikuti dengan tempat)

Berdasarkan teori Chomsky (1981), bentuk kata kerja ini digambarkan dalam bentuk kerangka agihan atau kerangka subkategori.

i. hurai : KK, [… FN]

ii. tukar : KK, [… FN,FN / FN FP]

iii. jalan : KK, […]

iv. adalah : KK, [… FN]

v. ialah : KK, [… FN]

vi. merupakan : KK, [… FN]

Menurut Ramli (2006), bentuk kerangka agihan di atas boleh ditukar dalam bentuk struktur argumen yang mewakili jenis subjek, kata kerja dan predikat.

i. hurai : KK, 1 2
FN FN

ii. tukar : KK, 1 2 3
FN FN FN
FN FN FP

iii. jalan : KK, 1
FN

iv. adalah : KK, 1 2
FN FN

v. ialah : KK, 1 2
FN FN

vi. merupakan : KK, 1 2
FN FN

2.10.2 Petua Peranan Tematik

Peranan teta atau peranan tematik merujuk kepada hubungan semantik antara kata kerja dan argumennya (Ramli 2006). Sebagai contoh, kata kerja menghurai memerlukan dua argumen yang ditandakan dengan peranan tematiknya. Argumen

subjek ditandakan sebagai Pelaku manakala argumen objek pula ditandakan sebagai Penderita.

a Jenis kata kerja dalam penandaan peranan tematik Bahasa Melayu

Ramli (2006) telah menetapkan tujuh peranan tematik signifikan yang ditandakan oleh kata kerja iaitu Pelaku, Penderita, Tema, Tempat, Pemanfaat dan Pengalami. Sri Liyaningsih & Siti Zuhriah Ariatmi (2016) pula menyatakan terdapat 12 peranan tematik yang ditandakan oleh kata kerja kepada argumennya (tambahan Alatan, Bilangan, Tujuan, Alasan dan Masa). Walaubagaimanapun, bilangan dan jenis peranan tematik yang dipilih adalah berdasarkan kepada set data latihan dan ujian dalam sesebuah kajian.

Kesemua peranan tematik tersebut boleh dihuraikan seperti contoh-contah ayat berikut:

i. Penganalisis sintaksis menjelajah token.

PELAKU KK PENDERITA

ii. Pengkompil merupakan penterjemah.

TEMA KK TEMA

iii. Nahu bebas konteks mewakili peraturan sintaks sesuatu bahasa.

TEMA KK. PENGALAMI

iv. Nahu bebas konteks mempunyai terminal sebagai perwakilan token.

TEMA KK TEMA

dalam bahasa.

TEMPAT

v. Pengkompil menukar arahan daripada bahasa paras tinggi kepada

PELAKU KK. PENDERITA SUMBER

bahasa mesin.

PEMANFAAT

Penandaan peranan tematik dianggap berlaku pada tahap struktur dasar. Ini bemakna peranan itu kekal walaupun ayat aktif diubah kepada ayat pasif. Sebagai contoh, frasa nama ‘Penganalisis sintaksis’ dalam ayat (i) ‘Penganalisis sintaksis menjelajah token’ akan kekal sebagai Pelaku walaupun struktur argumen ditukar kepada ‘Token dijelajah oleh penganalisis sintaksis’.

b Imbuhan dalam penandaan peranan tematik Bahasa Melayu

Sebelum ini telah dinyatakan bahawa penandaan tematik Bahasa Melayu ditentukan berdasarkan predikat atau kata kerja kepada argumennya samada argumen luar iaitu frasa nama subjek atau argumen dalam iaitu frasa nama objek (langsung atau tak langsung) (Chomsky 1981). Ayat (i) hingga (v) merupakan contoh bagaimana peranan tematik ditanda berdasarkan jenis kata kerja samada transitif, dwitransitif dan tak transitif. Namun begitu, dalam Bahasa Melayu, selain dari faktor jenis kata kerja, imbuhan turut mempengaruhi dalam penandaan peranan tematik ini. Sebagai contoh:

i. hurai

a. Penghurai <u>meng</u>hurai ayat.

b. Penghurai <u>meng</u>hurai<u>kan</u> ayat kepada token.

ii. lahu

a. Pemproses sedang <u>me</u>lahu.

b. Aturcara <u>me</u>lahu<u>kan</u> pengkompil itu.

c. Aturcara <u>me</u>lahu<u>kan</u> pengkompil itu di peringkat proses

Dalam contoh-contoh ayat di atas menunjukkan struktur argumen yang berbeza bagi kata kerja ‘hurai’ dan ‘lahu’. Perbezaan tersebut dipengaruhi oleh elemen imbuhan awalan dan akhiran yang diimbuhkan ke atas kata kerja tersebut.

i. hurai

a. menghurai:	KK;	1	2	
		FN	FN	
b. menghuraikan:	KK;	1	2	3
		FN	FN	FN

ii. lahu

a. melahu:	KK;	1		
		FN		
b. melahukan:	KK;	1	2	
		FN	FN	
c. melahukan:	KK;	1	2	(3)
		FN	FN	FN

Kata kerja ‘hurai’ merupakan kata kerja transitif manakala kata kerja ‘lahu’ ialah kata kerja tak transitif. Walaubagaimanapun, struktur argumen ini berubah dipengaruhi oleh penambahan imbuhan ke atas kata kerja tersebut. Secara tidak langsung, imbuhan ini memberi kesan signifikan ke atas peranan tematik yang dimainkan oleh kata kerja dalam ayat.

i. hurai

a. Penghurai menghurai ayat.

PELAKU PENDERITA

b. Penghurai menghuraikan ayat kepada token.

PELAKU PENDERITA PEMANFAAT

ii. lahu

a. Pemproses sedang melahu.

PENGALAMI

b. Aturcara melahukan pengkompil itu.

PELAKU PENDERITA

c. Aturcara melahukan pengkompil itu di peringkat proses

PELAKU PENDERITA TEMPAT

Daripada perbincangan ini, jelas menunjukkan peranan tematik Bahasa Melayu bukan sahaja dipengaruhi oleh jenis kata kerja itu sendiri (samada transitif atau tak transitif), namun penambahan imbuhan awalan dan akhiran juga memberi kesan signifikan ke atas struktur argumen ayat dan memberi kesan ke atas Petua Peranan Tematik Bahasa Melayu.

2.11 PANGKALAN DATA LEKSIKAL DAN KAEDAH RANGKAIAN SEMANTIK

Padanan kata atau frasa dibuat berdasarkan maklumat rangkaian semantik dalam pangkalan data leksikal *Wordnet*.

2.11.1 Pangkalan Data Leksikal

Leksikal adalah berkaitan dengan kata atau perbendaharaan kata bagi sesebuah bahasa. Unit leksikal merupakan kata tunggal, sebahagian daripada perkataan atau

rantaian perkataan yang membentuk elemen asas leksikon bahasa, iaitu perbendaharaan kata.

Pangkalan data leksikal pula menyimpan maklumat leksikal bagi setiap perkataan. Maklumat leksikal terdiri daripada kategori leksikal dan perkataan sinonim, termasuklah hubungan semantik dan fonologi antara perkataan atau set perkataan. Bagi kajian ini, pangkalan data leksikal merangkumi kategori leksikal kata nama, kata kerja, kata adjektif dan kata sendi.

Pangkalan data leksikal berbeza daripada kamus. Ianya mengambilkira makna (*sense*) kata, sebaliknya kamus hanya mengambilkira senarai kata. Dengan kata lain, pangkalan data leksikal menjurus kepada pemahaman konsep ayat dengan mengekstrak hubungan semantik setiap *synset* (token/kata) yang signifikan dalam sesebuah ayat.

2.11.2 Hubungan Dan Rangkaian Semantik

Semantik dengan definisi ringkas ialah makna. Namun untuk memahami makna sebenar sesebuah ayat, perlu mendalami makna sebenar setiap perkataan yang dibina terlebih dahulu. Makna sebenar setiap perkataan tersebut pula diperolehi dengan memetakan hubungan semantik setiap satunya dan diproses sehingga difahami.

Hubungan semantik setiap perkataan dalam ayat diukur dan diproses menggunakan rangkaian semantik. Rangkaian semantik akan mengumpulkan perkataan kepada set sinonim yang dikenali sebagai *synset*. Ianya juga menyediakan definisi ringkas, contoh penggunaan dan merekodkan bilangan hubungan antara set-set sinonim tersebut atau ahli-ahlinya. Kata kerja, kata nama, kata adjektif dan kata sendi akan dikumpulkan pada set sinonim kognitif, setiap satunya menunjukkan konsep yang berbeza. *Synset* akan disaling-hubungkan antara konsep semantik dan hubungan leksikal. Hasil rangkaian semantik ini digunakan sebagai salah satu kaedah pengukuran kesetaraan ayat dengan membuat perbandingan antara ayat dari segi konsep, bukan dari segi makna langsung.

2.11.3 Kaedah Pengukuran Rangkaian Semantik

Pengukuran kesetaraan rangkaian semantik menggunakan maklumat yang ditemui dalam hierarki konsep ‘ialah’ (*is-a*), dan mengira kadar persamaan antara konsep A dan konsep B (Petersen et. al 2004). Sebagai contoh, pengukuran kesetaraan akan menunjukkan ‘epal’ lebih setara dengan ‘anggur’ berbanding ‘kereta’, berdasarkan fakta di mana ‘epal’ dan ‘anggur’ berkongsi ‘buah’ sebagai *ancestor* dalam hierarki kata nama. Walaubagaimanapun, Noor Syakirah et. al. (2011) menyatakan bahawa hubungan semantik dalam senibina *Wordnet* tidak hanya terhad kepada hubungan ‘ialah’ untuk ‘sinonim’, namun terdapat juga hubungan semantik ‘antonim’, ‘hiponim’ dan ‘hipernim’, ‘meronim’, ‘holonim’ dan ‘troponim’ (nama sifat). Namun begitu, hubungan ‘ialah’ adalah hubungan semantik yang paling banyak digunakan dalam *Wordnet* (Thabet Slimani 2013).

a. *Path length*

Kaedah pengukuran laluan pengiraan-nod yang ringkas. Markah kerelatifan adalah berkadar songsang dengan bilangan nod sepanjang laluan terdekat antara *synset*. Kebarangkalian laluan terdekat terjadi apabila dua *synset* adalah sama, di mana panjang laluan tersebut bernilai 1. Dengan itu, nilai kerelatifan maksimum adalah 1.

b. *Leacock & Chodorow*

Pengukuran kerelatifan yang diperkenalkan oleh *Leacock & Chodorow* (*lch*) ialah $\log(panjang/(2\times D))$, di mana panjang adalah panjang laluan terpendek antara dua *synset* (menggunakan pengiraan-nod) dan *D* adalah kedalaman maksimum taksonomi.

Pengukuran *lch* mengambilkira kedalaman taksonomi di mana *synset* yang ditemui sangat dipengaruhi oleh kehadiran nod akar unik. Jika terdapat nod akar unik, ini bermakna terdapat hanya dua taksonomi; satu untuk kata nama dan satu untuk kata kerja. Kesemua kata nama dan kata kerja seterusnya akan dikelompokkan ke dalam taksonomi masing-masing.

Sebaliknya, jika nod akar tidak digunakan, ini bermakna ada kemungkinan bagi sesebuah *synset* dimiliki oleh lebih dari satu taksonomi. Dalam kes ini, kerelatifan diukur dengan mengira laluan terdekat antara *synset* menggunakan rumus *Least Common Subsummer (LCS)*. Nilai *D* adalah kedalaman maksimum taksonomi tersebut di mana *LCS* ditemui. Jika *LCS* dimiliki oleh lebih dari satu taksonomi, maka taksonomi dengan nilai kedalaman maksimum tertinggi akan dipilih.

c. *Wu & Palmer*

Pengukuran *Wu & Palmer* (wup) mengira kerelatifan dengan mengambilkira kedalaman dua *synset* dalam taksonomi rangkaian semantik, bersama-sama dengan kedalaman *LCS*. Rumus yang digunakan adalah $2 \times \frac{kedalaman(lcs)^2}{kadalaman(S_1) + kedalaman(S_2)}$.

Ini bermakna julat nilai adalah $0 \leq nilai \leq 1$. Nilai tersebut tidak mungkin bernilai 0 kerana kedalaman *LCS* tidak akan bernilai 0. Nilai 1 jika dua *synset* adalah sama dari segi semantiknya.

d. *Resnik*

Nilai relatif adalah sama dengan kandungan maklumat (*IC*) bagi *LCS*. Ini bermakna nilai yang diperolehi adalah sentiasa lebih besar atau sama dengan sifar. *Upper bound* bagi nilai tersebut secara umumnya agak tinggi dan berbeza-beza bergantung kepada saiz korpus yang digunakan untuk mengenalpasti nilai kandungan maklumat. Lebih tepat lagi, *upper bound* sepatutnya *ln(N)* di mana *N* adalah bilangan kata dalam korpus.

e. *Jiang & Conrath*

Nilai kerelatifan diukur menggunakan pengukuran *Jiang & Conrath* (*jcn*) iaitu $\frac{1}{jarak_{jcn}}$, di mana $jarak_{jcn}$ bersamaan dengan $IC(synset_1) + IC(synset_2) - 2 \times IC(lcs)$.

Terdapat dua kes khas yang perlu diberi perhatian apabila pengukuran kerelatifan dibuat; kedua-dua kes tersebut melibatkan $jarak_{jcn}$ bersamaan dengan sifar. Dalam kes pertama, $IC(synset_1) = IC(synset_2) = IC(lcs) = 0$. Dalam keadaan sebenar, kes ini hanya berlaku apabila ketiga-tiga *synset₁*, *synset₂* dan *lcs* adalah nod akar. Walaubagaimanapun, apabila pengiraan kekerapan *synset* adalah sifar, kita akan menggunakan nilai 0 untuk kandungan maklumat. Dalam kes ini, kita akan mengembalikan nilai 0 disebabkan oleh kekurangan data.

Dalam kes kedua, situasinya adalah $IC(synset_1) + IC(synset_2) - 2 \times IC(lcs)$. Kebiasaannya kes ini ditemui berlaku apabila $IC(synset_1) = IC(synset_2) = IC(lcs)$ (contohnya kedua-dua input *synset* adalah sama). Secara intuitifnya, kes ini berlaku apabila ianya mencapai kerelatifan maksimum, di mana bernilai infiniti, namun mustahil untuk memulangkan nilai infiniti. Sebaliknya, kita akan memperolehi kemungkinan jarak terdekat yang lebih besar dari sifar dan memulangkan hasil pendaraban songsang bagi jarak tersebut.

f. *Lin*

Nilai kerelatifan berdasarkan pengukuran Lin adalah bersamaan dengan, $\frac{2 \times IC(lcs)}{IC(synset_1) + IC(synset_2)}$ di mana *IC(x)* adalah kandungan maklumat bg *x*. Julat nilai kerelatifan adalah lebih besar atau sama dengan sifar dan lebih kecil atau sama dengan satu.

Jika kandungan maklumat dari salah satu *synset₁* atau *synsets₂* adalah sifar, maka nilai kerelatifan yang dipulangkan adalah 0, disebabkan oleh kekurangan data. Kesimpulannya, kandungan maklumat bagi sesebuah *synset* adalah sifar hanya jika *synset* tersebut merupakan nod akar, namun apabila kekerapan *synset* adalah 0, kita akan menggunakan nilai sifar sebagai kandungan maklumat disebabkan oleh kekurangan pilihan.

g. *Hirst & St-Onge*

Pengukuran ini *Hirst & St-Onge* (hso) berfungsi dengan mencari rantaian leksikal yang menghubungkan dua makna perkataan (*word sense*). Terdapat tiga kelas hubungan yang dipertimbangkan: lebih relatif, relatif, dan sederhana relatif. Nilai kerelatifan maksima ialah 16.

2.12 STRUKTUR AYAT BAHASA MELAYU

Setiap ayat dibina berdasarkan struktur ayat yang spesifk. Walaupun secara umumnya setiap ayat mempunyai subjek dan predikat, namun faktor jenis bahasa memainkan peranan dari segi susunan nahu tersebut. Bagi kajian ini yang memfokuskan Bahasa Melayu, maka perbincangan bab ini hanya berkisar tentang struktur ayat Bahasa Melayu sahaja.

2.12.1 Definisi Ayat

Menurut Nik Safiah (1995), ayat ditakrifkan dengan ringkasnya sebagai ucapan yang bermakna. Ucapan tersebut dibina dari gabungan kata menurut petua sintaksis yang betul untuk menghasilkan ayat Bahasa Melayu yang gramatis. Petua sintaksis dinyatakan sebagai bidang ilmu bahasa yang mengkaji proses pembinaan ayat (Nik Safiah et. al 2010). Proses pembinaan ayat mesti menepati kaedah penggabungan dan urutan perkataan tunggal atau perkataan berkelompok berdasarkan hukum-hukum tertentu dalam Bahasa Melayu. Dengan kata lain, ayat dibina dari rangkaian kata dengan kata yang lain dan membentuk frasa, lalu frasa digabungkan dengan frasa lain lalu akhirnya menjadi sebuah ayat.

2.12.2 Golongan Kata

Kata mempunyai pelbagai makna bergantung kepada penggunaannya dalam sesebuah ayat. Berdasarkan pada sintaks Bahasa Melayu, kata digolongkan kepada empat kategori iaitu:

i. Kata Nama

ii. Kata Kerja

iii. Kata Adjektif

iv. Kata Tugas

i Kata nama

Dari segi makna, kata nama adalah kata yang merujuk kepada sesuatu, baik yang bernyawa atau tak bernyawa, konkrit atau abstrak (Asmah 2009).

a. Kata nama khas

Kata nama khas ialah kata nama yang digunakan untuk orang, haiwan, tempat, benda, institusi, pangkat dan yang seumpamanya secara khusus. Dari segi petua penulisan pula, setiap perkataan untuk kata nama khas mesti bermula dengan huruf besar. Contohnya,

Universiti Kebangsaan Malaysia
Ular Sawa

b. Kata nama am

Kata nama am pula digunakan untuk objek yang sama seperti kata nama khas, namun bersifat umum. Contohnya,

universiti
ular

c. Kata Ganti Nama

Kata ganti nama merujuk kepada kata yang menggantikan kata nama khas dan kata nama am.

ii Kata kerja

Kata kerja adalah kata yang bersifat dengan perbuatan atau keadaan melakukan sesuatu. Kata kerja boleh dikelaskan kepada dua bahagian:

a. Kata kerja tak transitif

Kata kerja yang tidak memerlukan penyambut atau objek sesudahnya (Nik Safiah et. al 2010). Contohnya,

Adik sedang makan.

b. Kata kerja transitif

Kata kerja yang tidak boleh berdiri sendiri, iaitu memerlukan penyambut atau objek sesudahnya. Contohnya,

Ali menyepak bola itu
Bola itu disepak oleh Ali.

iii Kata adjektif

Kata kerja yang tidak boleh berdiri sendiri, iaitu memerlukan penyambut atau objek sesudahnya. Contohnya,

Anak guru itu amat baik.
Pegawai polis itu sangat tegas.

iv Kata tugas

Kata Tugas adalah satu perkataan yang hadir dalam ayat, klausa atau frasa untuk mendukung sesuatu tugas sintaksis tertentu, sama ada sebagai penghubung, penerang, penentu, penguat, pendepan, pembantu, penegas, penafi, pembenar, pemeri atau tugas-tugas lain. Contohnya,

Siti dan Aminah adalah adik-beradik.
Kereta itu sangat cantik

Dalam seksyen ini perbincangan kajian terdahulu akan dibahaskan. Perbahasan akan ditekankan ke atas kerangka kerja atau teknik, data penilaian, jenis penyelesaian dan hasil penilaian yang didapati bagi setiap kajian terdahulu.

2.12.3 Frasa Dalam Binaan Ayat Bahasa Melayu

Frasa merupakan suatu unit kata yang berpotensi dikembangkan atau diperluas menjadi dua atau lebih perkataan yang memberi makna (Nabillah Bolhassan & Che Ibrahim 2014). Dengan kata lain, frasa adalah gabungan dua perkataan atau lebih. Konstituen frasa dibina dari gabungan inti dan penerang. Inti ditakrikfkan sebagai unsur yang diberi penumpuan makna dan mewakili seluruh frasa dari segi makna, sementara penerang merupakan unsur yang menerangkan makna dalam inti itu.

i Frasa nama

Frasa nama ialah binaan yang terdiri samada daripada satu atau lebih perkataan yang berfungsi sebagai satu konstituen dalam binaan ayat. Konstituen frasa nama mengandungi binaan inti + penerang dan inti + inti. (Nik Safiah et. al 2010).

a. inti + penerang

- inti + penerang nama
- inti + penerang bukan nama

b. inti + inti

- inti + inti dengan maksud sama erti
- inti + inti dengan maksud lawan erti

Bagi binaan inti + penerang nama, penerang nama merangkumi nama keturunan, nama jenis, nama penyambut, nama kegunaan, nama kelamin, nama tempat, nama arah, nama anggota badan, nama tenaga penggerak, nama perihal, nama

milik, nama khas dan nama gelaran. Penerang bukan nama pula merangkumi penentu, kata kerja, kata adjektif, kata adverb, frasa sendi nama dan bilangan ordinal.

ii Frasa kerja

Frasa kerja adalah binaan yang boleh terdiri daripada satu atau lebih perkataan dan kata intinya kata kerja. Kata kerja boleh dibahagikan kepada kata kerja tak transitif atau kata kerja transitif. Kata kerja tak transitif boleh hadir sendirian sebaliknya kata kerja transitif mesti diikuti oleh objek.

a. Frasa kerja tak transitif

Frasa kerja tak transitif tidak memerlukan objek untuk melengkapkannya. Terdapat dua kategori kata kerja tak transitif iaitu kata kerja tak transitif tanpa pelengkap dan kata kerja tak transitif dengan pelengkap.

b. Frasa kerja transitif

Frasa kerja transitif pula memerlukan frasa nama sebagai objek untuk melengkapkannya. Frasa kerja transitif ini boleh dibahagikan kepada dua kategori iaitu kata kerja yang bertindak sebagai ayat pasif dan kata kerja yang bertindak sebagai ayat aktif (Nik Safiah 1995). Bagi setiap kategori itu pula, frasa kerjanya boleh terdiri daripada samada satu atau dua objek (Nik Safiah et. al 2010).

iii Frasa adjektif

Frasa adjektif merupakan sebuah susunan perkataan (rangkai kata) di mana intinya terdiri daripada kata adjektif atau kata sifat. Frasa adjektif yang dibincangkan ialah yang berfungsi sebagai predikat dalam ayat. Contohnya,

Ujian memandu itu *sangat berjaya*.
Hari ini *panas sungguh*.

iv Frasa sendi nama

Frasa sendi nama ialah sebuah susunan perkataan yang mana intinya terdiri daripada kata sendi nama diikuti dengan frasa nama. Contohnya,

di universiti
sejak hari semalam

2.12.4 Struktur Binaan Ayat

Sesebuah ayat terbentuk dari gabungan dua konstituen; subjek dan predikat (Nik Safiah 1995). Subjek adalah konstituen bagi perkara/topik yang dibincangkan manakala predikat merupakan konstituen yang membincangkan tentang subjek tersebut. Jadual 2.4 menunjukkan contoh-contoh ayat yang dibahagikan kepada subjek dan predikat.

Jadual 2.4 Subjek dan predikat Bahasa Melayu

Ayat	
Subjek	**Predikat**
Dia	ketua kelas 1 Pintar.
Pokok itu	ditebang oleh Samad.
Lariannya	sungguh laju.
Batalion aksar itu	akan ke medan perang.

2.12.5 Corak Asas Ayat

Menurut Nik Safiah et. al (2010), terdapat empat corak asas pembentukan sesebuah ayat Bahasa Melayu. Kesemua corak tersebut adalah dibangunkan daripada terbitan konstituen predikat. Kesemua subjek dibina daripada frasa nama (FN), namun konstituen predikat boleh dibina daripada mana-mana empat frasa tersebut; frasa nama (FN), frasa kerja (FK), frasa adjektif (FA) dan frasa sendi nama (FS). Corak-corak asas pembentukan ayat Bahasa Melayu beserta contoh dirumuskan dalam Jadual 2.5.

Jadual 2.5 Corak asas Bahasa Melayu

Corak	Subjek	Predikat
FN + FN	Dia	ketua kelas 1 Pintar.
FN + FK	Pokok itu	ditebang oleh Samad.
FN + FA	Lariannya	sungguh laju.
FN + FS	Batalion aksar itu	akan ke medan perang.

2.12.6 Ragam Ayat

Pembentukan ayat berdasarkan empat corak dalam Jadual 2.3. Dalam Bahasa Melayu, terdapat empat ragam ayat yang dikelaskan sebagai ayat penyata, ayat tanya, ayat perintah dan ayat seruan. Walaubagaimanapun, kajian ini hanya mengambilkira tiga jenis ragam ayat iaitu ayat tanya dan ayat perintah (sebagai soalan) dan ayat penyata (sebagai jawapan).

i Ayat tanya

Ayat tanya biasanya digunakan untuk menanyakan sesuatu hal atau perkara. Sebagai contoh,

Apakah tugas penghurai? (2.1)

Ayat (2.1) merupakan contoh ayat tanya dengan kehadiran kata tanya ‘apa’ dengan imbuhan ‘kan’ di awal ayat dan tanda soal ‘?’ di penghujung ayat. Antara kata tanya lain adalah seperti siapa, bila, di mana, bagaimana dan berapa.

ii Ayat perintah

Ayat perintah didefinisikan sebagai ayat yang digunakan untuk menghasilkan sesuatu tindakan. Dalam kajian ini, ianya digunakan sebagai ayat soalan dan jawapan pelajar adalah merupakan tindakan.

Nyatakan komponen pertama pengkompil. (2.2)

Ayat (2.2) dianggap sebagai ayat perintah dengan wujudnya kata perintah ‘nyata’ dengan imbuhan ‘kan’ di awal ayat. Ayat ini tidak memerlukan symbol ‘?’ untuk berfungsi sebagai sebuah soalan, memadai dengan kata perintah tersebut. Selain itu, kata ‘beri’, ‘nama’, ‘jelas’ dan ‘hurai’ adalah antara kata-kata perintah yang lain.

iii Ayat penyata

Sebaliknya, ayat penyata bertindak sebagai ayat yang menyatakan atau menerangkan sesuatu hal, dalam kes ini menjurus kepada menjawab soalan yang diberikan. Selain itu, ayat penyata juga dikenali sebagai ayat keterangan. Pembinaan sesebuah ayat penyata tidak memerlukan sebarang kehadiran perkataan spesifik, namun konteks ayat yang dibina haruslah relevan dengan soalan yang diutarakan.

2.12.7 Struktur Bahasa Melayu Dalam Pengukuran Kesetaraan Ayat

Jenis bahasa adalah faktor signifikan dalam pengukuran kesetaraan ayat. Berpandukan pada bentuk struktur unik sesebuah bahasa, kebanyakan algoritma pengukuran kesetaraan ayat dibina khusus untuk bahasa itu sahaja, lebih-lebih lagi apabila ianya mengaplikasikan teknik-teknik PBT berbanding teknik-teknik statistik.

Menurut Nik Safiah (1995) lagi, Bahasa Melayu secara umumnya boleh dibahagikan kepada dua kategori; [1] ayat ringkas dan [2] ayat majmuk. Ayat ringkas dibina dengan satu subjek dan satu predikat. Contohnya,

Penghurai menghurai token. (2.3)

Ayat (2.3) adalah contoh ayat ringkas yang mana mewakili kewujudan satu subjek (Penghurai) dan satu predikat (menghurai token). Predikat ini pula terdiri daripada satu kata kerja (menghurai) dan satu objek (token). Manakala, ayat kompleks pula boleh dibina daripada lebih dari satu subjek atau predikat. Dalam BM, terdapat tiga jenis ayat kompleks iaitu ayat majmuk gabungan, ayat majmuk pancangan dan ayat majmuk campuran.

Tiga komponen utama dalam pengkompil adalah leksikal, sintaksis dan semantik. (2.4)

Apabila sesuatu token itu wujud, pengkompil akan memulangkan nilai true. (2.5)
Nahu bebas konteks boleh dihurai lagi kerana mempunyai kes-kes sensitif seperti ambiguiti dan rekursif kiri. (2.6)

Ayat majmuk gabungan adalah suatu keadaan di mana dua atau lebih ayat digabungkan bersama menggunakan kata sendi seperti ‘dan’, ‘atau’ dan ‘tetapi’. Sebagai contoh, ayat (2.4) boleh dipecahkan kepada subjek (komponen utama, pengkompil), kata kerja (adalah) dan objek (leksikal, sintaksis, semantik). Kata sendi ‘dan’ digunakan untuk menghubungkan kesemua tiga jenis objek yang sama. Ayat majmuk pancangan adalah suatu keadaan di mana satu atau lebih ayat atau klausa bebas dipancangkan ke dalam ayat atau klausa utama. Ayat (2.5) merupakan contoh ayat majmuk pancangan di mana ayat ‘sesuatu token itu wujud’ dipancangkan kepada ayat utama iaitu ‘pengkompil akan memulangkan nilai true’ menggunakan kata hubung ‘apabila’. Ayat majmuk campuran pula terbentuk dari gabungan ayat tunggal dan ayat majmuk. Berdasarkan ayat (2.6), kehadiran kata sendi ‘kerana’ telah menghubungkan ayat tunggal ‘Nahu bebas konteks boleh dihurai lagi’ dan ayat majmuk ‘mempunyai kes-kes sensitif seperti ambiguiti dan rekursif kiri’. Ayat kedua merupakan ayat majmuk dengan wujudnya kata sendi ‘dan’ dalam ayat tersebut.

2.12.8 Pengukuran Kesetaraan Ayat

Pengukuran kesetaraan ayat ditentukan dengan pemadanan Petua Peranan Tematik, kesetaraan *synset* dan kesetaraan ayat. Setelah itu, markah akan dikira dari segi bilangan poin dan markah yang diperuntukkan untuk setiap soalan.

2.13 RUMUSAN

Bab ini pada awalnya membincangkan berkenaan PEPB dan perkembangan pencapaiannya sehingga kini. Kemudian empat jenis kaedah yang digunakan untuk mengukur kesetaraan ayat dalam PEPB iaitu kaedah statistik, kaedah semantik, kaedah hibrid dan kaedah linguistik, diperjelaskan. Kaedah-kaedah tersebut dihuraikan dari segi pencapaian kajian dalam mengukur kesetaraan ayat supaya

ketepatan ukurannya menghampiri penilaian manusia. Hasil dari kajian ke atas kaedah-kaedah tersebut, didapati kaedah linguistik, iaitu Petua Peranan Tematik, mampu menghasilkan pemadanan struktur argumen yang lebih baik. Petua ini membuat penandaan peranan yang dimainkan oleh setiap argumen yang signifikan dan menyediakan maklumat berkenaan hubungan antara argumen dan peranan yang dimainkan dalam menentukan konteks argumen tersebut dalam sesebuah ayat. Manakala, bagi mengukur kesetaraan ayat, Teknik Rangkaian Semantik yang mengira kadar kesetaraan *synset* dan kesetaraan ayat menggunakan kaedah wup dijangka mampu memberikan ketepatan yang tinggi berbanding penilaian manusia. Di akhir bab, isu-isu yang melibatkan pangkalan data leksikal dan struktur ayat Bahasa Melayu diperincikan.

BAB III

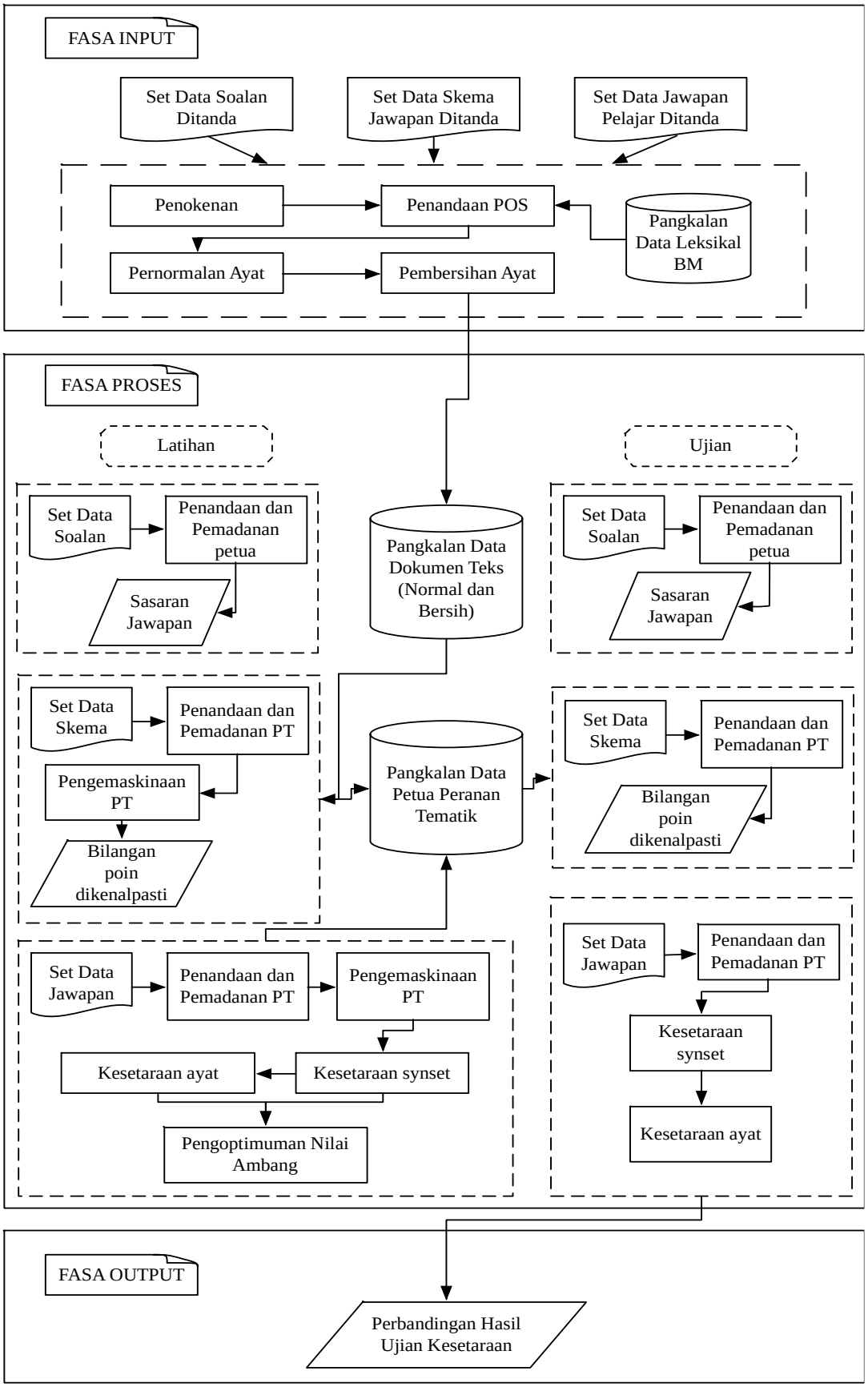
METODOLOGI

3.1 PENGENALAN

Penilaian Esei Pendek Bersepadu (PEPB) dalam kajian ini menilai esei pendek Bahasa Melayu pelajar dengan cara mengukur kesetaraan ayat melalui gabungan kaedah linguistik iaitu Petua Peranan Tematik dan teknik statistikal iaitu Teknik Rangkaian Semantik. Petua Peranan Tematik bertujuan untuk mengenalpasti peranan yang dimainkan oleh argumen subjek dan objek berdasarkan kata kerja yang wujud dalam sesebuah ayat. Manakala, Teknik Rangkaian Semantik pula berfungsi untuk mengira nilai relatif kesetaraan hubungan sematik antara dua *synset* (argumen subjek, kata kerja dan argumen objek). Kedua-dua kaedah ini merupakan proses utama setelah input iaitu jawapan pelajar dalam bentuk teks (ayat), dimasukkan. Sebagai hasil akhir, markah bagi setiap set jawapan tersebut diperolehi dengan mengukur kesetaraan ayat antara jawapan pelajar dan skema jawapan berdasarkan bilangan poin yang diperuntukkan.

3.2 SENIBINA PROSES PENGUKURAN KESETARAAN AYAT

Rajah 3.1 menunjukkan kaedah kajian yang dijalankan. Berdasarkan rajah tersebut, keseluruhan proses pengukuran kesetaraan ayat antara jawapan pelajar dan skema jawapan terbahagi kepada tiga fasa iaitu Fasa Input, Fasa Proses dan Fasa Output. Setiap fasa tersebut mengandungi sub-proses dan sebahagiannya menghasilkan output yang menjadi input kepada sub-proses yang lain.



Rajah 3.1 Senibina proses pengukuran kesetaraan.

3.3 FASA INPUT

Fasa ini merupakan fasa pra-proses di mana input-input yang diterima akan menjalani beberapa proses asas iaitu penokenan dan penandaan golongan kata (POS) dan proses khusus iaitu penormalan dan pembersihan ayat. Terdapat tiga input yang diterima dalam fasa ini iaitu set soalan ditanda, set skema jawapan ditanda dan set jawapan ditanda. Set-set data ini telah ditanda oleh penilai manusia dari segi Petua Subjek-Predikat (kata nama + kata kerja + objek) dan markah bagi setiap jawapan pelajar bagi tujuan latihan dan ujian. Dengan ini, perbandingan antara markah jawapan pelajar yang dinilai oleh teknik yang digunakan dan markah yang dinilai oleh manusia dapat dilaksanakan.

3.3.1 Penokenan

Pada peringkat awal Fasa Input, ketiga-tiga jenis set data akan melalui proses penokenan. Penokenan adalah proses di mana sesebuah ayat akan dipecahkan kepada kata, nombor dan simbol, dipanggil token.

3.3.2 Penandaan Golongan Kata

Penanda golongan kata akan menanda jenis golongan kata bagi setiap token yang berkaitan. Setelah semua aksara dipecahkan dalam bentuk susunan token, token-token tersebut akan dirangkaikan untuk membentuk frasa-frasa tertentu berdasarkan dua petua iaitu petua kata nama khas dan petua kata majmuk. Petua kata nama khas dikenalpasti melalui kata nama berturutan yang mempunyai huruf besar pada permulaan kata dan petua kata majmuk diperolehi daripada pangkalan data kata majmuk. Dalam kajian ini, penandaan golongan kata dibuat berdasarkan kepada penanda golongan kata Bahasa Melayu Mi-POS yang telah dibangunkan oleh Muhamed Lubani (Xian et. al 2016). Penanda ini dibangunkan berpandukan pendekatan pembelajaran mesin iaitu Model *MaxEnt* yang menggunakan peralatan Pemprosesan Bahasa Tabii (PBT) iaitu kod terbuka *OpenNLP*. Sebanyak 28 jenis golongan kata yang ditanda menggunakan penanda ini. Bagaimanapun, terdapat empat

jenis golongan kata didapati memberi implikasi yang tinggi PEBP iaitu kata nama, kata kerja, kata adjektif dan kata tugas.

Bagi petua kata nama khas, padanan dibentuk berpandukan kepada huruf besar pertama setiap perkataan yang berturutan. Manakala bagi petua kata majmuk, ianya terdiri daripada empat jenis kata majmuk iaitu kata nama majmuk, kata kerja majmuk, kata adjektif majmuk dan kata tugas majmuk. Selain itu, terdapat juga kata majmuk istilah khusus yang sedia wujud dalam pangkalan data leksikal Bahasa Melayu. Sebagai contoh,

i. Pada analisis sintaks,ia akan menyemak samada aksara tersebut sah atau tidak dengan menyemak keseluruhan token pada analisis semantik benar dari bina ayat *Bahasa Melayu*.

ii. Penganalisis leksikal, sintaksis, semantik dan penjana kod pertengahan merupakan *komponen pengkompil*.

iii. Penganalisis sintaksis akan *mereka bentuk* bahasa mengikut binaan yang sah.

iv. Pengkompil merupakan elemen penting dalam *atur cara*.

Ayat (i) adalah contoh rangkai kata yang membentuk frasa nama khas iaitu ‘Bahasa Melayu’. Frasa ini dikenalpasti berdasarkan huruf besar ‘B’ sebagai huruf pertama dalam perkataan pertama dan huruf besar ‘M’ sebagai huruf besar pertama dalam perkataan kedua. Ayat (ii) pula merupakan contoh ayat yang mengandungi kata nama majmuk iaitu hasil gabungan perkataan ‘komponen’ (kata nama) dan ‘pengkompil’ (kata nama). Kewujudan kedua-dua kata nama yang berturutan ditanda sebagai frasa nama berpandukan petua kata nama majmuk. Perkataan ‘mereka bentuk’ dalam Ayat (iii) dirangkaikan menjada frasa kerja berpandukan petua kata kerja majmuk. Seterusnya, perkataan ‘atur cara’ dalam ayat (iv) pula dirangkaikan sebagai kata majmuk khusus di mana frasa berjenis ini telah dimasukkan dalam pangkalan data leksikal Bahasa Melayu dalam kajian ini.

3.3.3 Penormalan Ayat

Setelah penandaan ke atas perkataan dan frasa selesai, ayat-ayat daripada set skema jawapan dan set data jawapan pelajar akan melalui proses penormalan. Pernormalan ayat bertujuan untuk mempiawaikan struktur ayat Bahasa Melayu kepada ayat yang lengkap. Ini kerana, Petua Peranan Tematik hanya berfungsi dengan baik ianya apabila dilaksanakan ke atas struktur ayat yang lengkap. Kesemua input adalah terdiri daripada set soalan, set skema jawapan dan set jawapan pelajar yang telah ditanda dengan jenis golongan kata masing-masing. Petua penormalan ayat adalah seperti berikut:

i. Menggantikan kata ganti nama (‘tersebut’, ‘nya’) atau ketiadaan subjek atau predikat dengan predikat atau subjek daripada ayat dalam soalan atau ayat dalam jawapan sebelum.

- Soalan: Namakan proses yang menterjemah token.
- Jawapan: Proses tersebut adalah penganalisis sintaksis.
- Penormalan: Proses yang menterjemah token adalah penganalisis sintaksis.

```
Function normalKGNSP():
Input: Ayat bagi soalan ke-N dan jawapan ke-N
Output: Ayat diperlengkapkan dengan subjek atau predikat
Pecahkan subjek-predikat ayat ke-n
Kenalpasti kewujudan kata ‘tersebut’ atau ketiadaan subjek atau predikat dalam ayat ke-n
If wujud kata ‘tersebut’ atau ketiadaan subjek atau predikat dalam ayat ke-n
        Kenalpasti kewujudan ayat ke-(n-1)
        If tidak wujud ayat ke-(n-1)
                If wujud kata ‘tersebut’ dalam subjek atau ketiadaan subjek dalam ayat ke-n
                        Ambil predikat dari ayat dalam soalan ke-N
                End if
                If wujud kata ‘tersebut’ dalam predikat atau ketiadaan predikat dalam ayat ke-n
                        Ambil subjek dari ayat dalam soalan ke-N
                End if
                Simpan dalam normalAyatTSP-1
        Else
                If wujud kata ‘tersebut’ dalam subjek atau ketiadaan subjek dalam ayat ke-n
                        Ambil predikat dari ayat ke-(n-1)
                End if
```

bersambung…

...sambungan

```
                    If wujud kata 'tersebut' dalam predikat atau ketiadaan predikat
                    dalam ayat ke-n
                        Ambil subjek dari ayat ke-(n-1)
                    End if
                    Simpan dalam normalAyatTSP-2
            End If
    End If
```

Rajah 3.2 Algoritma perlengkapan ketiadaan subjek atau predikat dalam ayat jawapan pelajar.

Rajah 3.2 menunjukkan algoritma perlengkapan ketiadaan subjek atau predikat dalam ayat jawapan pelajar. Fungsi *normalKGNSP()* menerima dua input iaitu ayat bagi soalan yang ke-*N* dan jawapan pelajar yang ke-*N*. Manakala, output bagi algoritma ini adalah merupakan ayat yang diperlengkapkan dengan subjek atau predikat.

Pada peringkat awal, ayat daripada set jawapan pelajar akan dipecahkan kepada bahagian subjek dan predikat. Seterusnya, algoritma akan menyemak ayat samada wujud kata 'tersebut' atau ketiadaan subjek atau predikat. Jika salah satu syarat tersebut dipenuhi, maka proses perlengkapan ayat ditentukan bergantung pada dua elemen. Elemen pertama ialah kewujudan ayat sebelum (*n-1*) bagi ayat ke-*n*. Elemen seterusnya pula ialah posisi kata 'tersebut' berada di subjek atau predikat atau ketiadaan subjek atau predikat dalam ayat.

Bagi elemen pertama keadaan pertama, jika ayat jawapan ke-*n* tidak mempunyai ayat sebelum (*n-1*), maka penentuan penggantian subjek atau predikat dibuat. Jika kata 'tersebut' berada dalam subjek atau ketiadaan subjek dalam ayat, maka penggantian subjek akan dilaksanakan. Penggantian tersebut dibuat dengan pengambilan subjek dari soalan dan digabungkan dengan predikat dalam jawapan pelajar. Sebaliknya, jika kata 'tersebut' berada dalam predikat atau ketiadaan predikat dalam ayat, maka penggantian predikat akan dilaksanakan. Penggantian tersebut dibuat dengan pengambilan predikat dari soalan

dan digabungkan dengan subjek dalam jawapan pelajar dan disimpan dalam *normalAyatTSP-1*.

Manakala bagi elemen pertama keadaan kedua, jika ayat jawapan ke-*n* mempunyai ayat sebelum (*n-1*), maka penentuan penggantian subjek atau predikat dibuat. Jika kata 'tersebut' berada dalam subjek atau ketiadaan subjek dalam ayat, maka penggantian subjek akan dilaksanakan. Penggantian tersebut dibuat dengan pengambilan subjek dari ayat sebelum dan digabungkan dengan predikat dalam jawapan pelajar. Sebaliknya, jika kata 'tersebut' berada dalam predikat atau ketiadaan predikat dalam ayat, maka penggantian predikat akan dilaksanakan. Penggantian tersebut dibuat dengan pengambilan predikat dari ayat sebelum dan digabungkan dengan subjek dalam jawapan pelajar dan disimpan dalam *normalAyatTSP-2*.

Sebagai hasil akhir, kedua-dua *normalAyatTSP-1* dan *normalAyatTSP-2* akan dihuraikan kepada struktur argumen bersesuaian. Struktur argumen yang mempunyai padanan tertinggi dari segi peranan tematik dan rangkaian semantik dipilih untuk pengiraan kesetaraan ayat.

ii. Pengstrukturan semula ayat yang dimulai dengan kata kerja

- Jawapan: Menghurai token merupakan tugas penghurai.
- Penormalan: Tugas penghurai merupakan menghurai token.

```
Function normalMulaKK():
Input: Ayat bagi jawapan pelajar ke-n
Output: Ayat distruktur semula
Kenalpasti token ke-1 dalam ayat bagi jawapan pelajar ke-n
If token ke-1 bersamaan dengan kata kerja
        Pecahkan subjek-predikat ayat ke-n berdasarkan kata pemeri
        Tandakan predikat, kata pemeri dan subjek dalam ayat
        Penstrukturan semula – subjek + kata pemeri + predikat
        Simpan dalam normalAyatMulaKK
End if
```

Rajah 3.3 Algoritma penstrukturan semula ayat.

Rajah 3.3 menunjukkan algoritma penstrukturan semula ayat yang dimulai dengan kata kerja. Fungsi *normalMulaKK()* menerima satu input iaitu ayat bagi jawapan pelajar dan satu ouput iaitu ayat yang telah distruktur semula.

Algoritma ini dimulakan dengan mengenalpasti jenis golongan kata bagi token pertama dalam ayat. Jika ianya berjenis kata kerja, maka ayat akan dipecahkan kepada bahagian subjek dan predikat berdasarkan kata pemeri ('ialah', 'adalah' dan 'merupakan') yang wujud. Proses pemecahan ayat tersebut akan menanda bahagian predikat, kata pemeri dan subjek. Akhir sekali, ayat tersebut akan struktur semula berdasarkan petua subjek + kata pemeri + predikat. Ayat yang telah distruktur semula tersebut disimpan dalam *normalAyatMulaKK*.

iii. Pemecahan ayat yang disambung dengan kata hubung

- Jawapan: *Penganalisis leksikal merupakan komponen utama pengkompil sementara penjana kod pertengahan merupakan komponen pengkompil yang biasa.*
- Penormalan:
  1. *Penganalisis leksikal merupakan komponen utama pengkompil.*
  2. *Penjana kod pertengahan merupakan komponen pengkompil yang biasa.*

```
Function normalKHPancangan():
Input: Ayat bagi skema jawapan ke-N dan jawapan pelajar ke-N
Output: Ayat dipecahkan berdasarkan kata hubung pancangan
Kenalpasti kewujudan kata hubung pancangan dalam ayat ke-n
If wujud kata hubung pancangan dalam ayat ke-n
        Loop bilangan kata hubung pancangan dalam ayat ke-n (i)
                Pecahkan ayat ke-n kepada sebelum dan selepas kata hubung
                pancangan
                Simpan dalam normalKHP[i]
        End loop
End if
```

Rajah 3.4 Algoritma pemecahan ayat berdasarkan kata hubung pancangan.

Rajah 3.4 menunjukkan algoritma pemecahan ayat berdasarkan kata hubung pancangan. Fungsi *normalKHPancangan()* menerima dua input iaitu ayat bagi skema jawapan ke-*N* dan jawapan pelajar ke-*N*.

Manakala output bagi fungsi ini merupakan ayat yang dipecahkan berdasarkan kata hubung pancangan namun diperlengkapkan berdasarkan petua pembentukan ayat subjek-predikat.

Di peringkat awal proses, jika didapati wujud kata hubung pancangan dalam ayat ke-*n*, maka ayat tersebut akan diproses dengan cara memecahkan kepada ayat yang berasingan. Bilangan penggelungan proses pemecahan ayat ini dibuat berdasarkan kepada bilangan kata hubung pancangan yang wujud dalam ayat. Penggelungan pemecahan ayat tersebut kepada sebelum dan selepas kata hubung pancangan dibuat sehingga kesemua ayat dipecahkan dan membentuk satu ayat lengkap berdasarkan petua subjek-predikat. Kemudiannya, kesemua ayat tersebut disimpan dalam *normalKHP[i]* (gelung ke-*i*, mewakili bilangan ayat yang dipecahkan berdasarkan bilangan kata hubung pancangan yang wujud dalam ayat ke-*n* untuk tujuan pengukuran kesetaraaan ayat.

iv. Pemecahan ayat yang disambung dengan simbol koma dan kata hubung gabungan (‘dan’, ‘atau’, ‘serta’)

- Jawapan: *Pengkompil melakukan analisis leksikal, sintaksis, semantik, menjana kod pertengahan dan menterjemah bahasa.*
- Penormalan:
  1. Pengkompil menganalisis leksikal
  2. Pengkompil menganalisis sintaksis
  3. Pengkompil menganalisis semantik
  4. Pengkompil menjana kod pertengahan
  5. Pengkompil menterjemah bahasa

```
Function normalKHGabungan():
Input: Ayat bagi jawapan skema jawapan ke-n dan jawapan pelajar ke-n
Output: Ayat dipecahkan berdasarkan kata hubung gabungan
Kenalpasti kewujudan simbol koma dan/atau kata hubung gabungan dalam
ayat ke-n
If wujud simbol koma dan/atau kata hubung gabungan dalam ayat ke-n
    Pecahkan ayat ke-n berdasarkan bilangan simbol koma dan kata
    hubung gabungan
    Simpan pecahan ayat ke-1 dalam normalKHG[0]
    Loop bilangan koma + kata hubung gabungan dalam ayat ke-n (i)
        Kenalpasti kewujudan subjek atau predikat dalam pecahan ayat
        ke-(i+1)
```

bersambung…

…sambungan

```
                    If pecahan ayat ke-(i+1) tidak mempunyai predikat daripada
                    ayat ke-1
                        Ambil predikat daripada ayat ke-1 dan gabungkan
                        dengan subjek pecahan ayat ke-(i+1)
                        Simpan dalam normalKHG[i+1]
                    End if
                    Else if pecahan ayat ke-(i+1) tidak mempunyai subjek daripada
                    ayat ke-1
                        Ambil subjek daripada ayat ke-1 dan gabungkan dengan
                        predikat pecahan ayat ke-(i+1)
                        Simpan dalam normalKHG[i+1]
                    End if
                End loop
            End if
```

Rajah 3.5 Algoritma pemecahan ayat berdasarkan simbol koma dan kata hubung gabungan.

Rajah 3.5 menunjukkan algoritma pemecahan ayat berdasarkan simbol koma dan kata hubung gabungan. Sepertimana fungsi *normalKHPancangan()*, fungsi *normalKHGabungan()* juga menerima dua input iaitu ayat bagi skema jawapan ke-*N* dan jawapan pelajar ke-*N*. Manakala output bagi fungsi ini merupakan ayat yang dipecahkan berdasarkan simbol koma dan kata hubung gabungan namun diperlengkapkan berdasarkan petua pembentukan ayat subjek-predikat.

Di peringkat awal proses, jika didapati wujud simbol koma dan/atau kata hubung gabungan dalam ayat ke-*n*, maka ayat tersebut akan diproses dengan cara memecahkan kepada ayat yang berasingan. Setelah itu, pecahan ayat ke-1 yang lengkap dari segi subjek dan predikatnya akan disimpan dalam *normalKHG[0]*. Seterusnya, penggelungan proses pemecahan ayat dimulakan. Bilangan penggelungan proses pemecahan ayat ini dibuat berdasarkan kepada bilangan simbol koma dan kata hubung gabungan yang wujud dalam ayat. Penggelungan pemecahan ayat tersebut kepada sebelum dan selepas simbol koma dan kata hubung gabungan dibuat sehingga kesemua ayat dipecahkan.

Jika didapati pecahan ayat ke-(*i+1*) tidak mengandungi predikat daripada pecahan ayat ke-1, maka predikat daripada ayat ke-1 tersebut akan diambil dan digabungkan dengan subjek pecahan ayat ke-(*i+1*) dan membentuk satu ayat lengkap berdasarkan petua subjek-predikat. Kemudiannya, ayat tersebut disimpan dalam *normalKHG[i+1]* (gelung ke-*i*, mewakili bilangan ayat yang dipecahkan berdasarkan bilangan simbol koma dan kata hubung gabungan yang wujud dalam ayat ke-*n*) untuk tujuan pengukuran kesetaraaan ayat.

Sebaliknya jika didapati pecahan ayat ke-(*i+1*) tidak mengandungi subjek daripada pecahan ayat ke-1, maka subjek daripada ayat ke-1 tersebut akan diambil dan digabungkan dengan predikat pecahan ayat ke-(*i+1*) dan membentuk satu ayat lengkap berdasarkan petua subjek-predikat. Kemudiannya, ayat tersebut disimpan dalam *normalKHG[i+1]* juga untuk tujuan sama iaitu pengukuran kesetaraaan ayat.

v. Penyengauan kata kerja ‘melakukan’ dengan kata kerja signifikan

- *melakukan analisis -> menganalisis*
- *melakukan proses analisis -> menganalisis*
- *melakukan imbasan –> imbasan*
- *melakukan pengecaman -> pengecaman*
- *melakukan pengujian -> pengujian*

```
Function normalKKMelakukan():
Input: Ayat bagi jawapan skema jawapan ke-N dan jawapan pelajar ke-N
Output: Ayat yang menyengaukan kata ‘melakukan’ + kata kerja signifikan
Kenalpasti kewujudan kesemua token dalam ayat
If wujud KK ‘melakukan’ dalam ayat
    If token selepas kata kerja ‘melakukan’ bersamaan dengan kata ‘proses’
        Buang kata ‘proses’
    End if
    Kenalpasti kata kerja signifikan selepas kata kerja ‘melakukan’
    Kenalpasti kata dasar bagi kata kerja signifikan
    If wujud imbuhan akhir dalam kata dasar
        Buang imbuhan akhir
    End if
    Kenalpasti aksara pertama dan kedua dalam kata dasar
    If kata dasar bermula dengan b, m, n, ny, ng, r, l, v, w atau y
```

bersambung…

…sambungan

```
16                  If kata dasar bermula dengan b atau v
17                          Imbuhan awal = ‘mem’
18                  End if
19                  Else
20                          Imbuhan awal = ‘me’
21                  End if
22          End if
23          Else if kata dasar bermula dengan p atau f
24                  Huruf pertama kata dasar digantikan dengan huruf ‘m’
25                  Imbuhan awal = ‘me’
26          End if
27          Else if kata dasar bermula dengan k, g, h, kh, gh, a, e i, o atau u
28                  If kata dasar bermula dengan k
29                          Huruf pertama kata dasar digugurkan
30                  End if
31                  Imbuhan awal = ‘meng’
32          End if
33          Else if kata dasar bermula dengan k
34                  Imbuhan awal = ‘meng’
35          End if
36          Else if kata dasar bermula dengan s
37                  Huruf pertama kata dasar digantikan dengan huruf ‘ny’
38                  Imbuhan awal = ‘me’
39          End if
40          Else if kata dasar terdiri daripada satu suku kata
41                  Imbuhan awal = ‘menge’
42          End if
43  End if
44  Gabungkan kata berdasarkan subjek + (imbuhan awal meN + kata kerja
    signifikan) + predikat dan disimpan dalam normalKKM
```

Rajah 3.6 Algoritma penyengauan kata kerja ‘melakukan’ dengan kata kerja yang lebih signifikan.

Rajah 3.6 menunjukkan algoritma penyengauan kata kerja ‘melakukan’ dengan kata kerja yang lebih signifikan yang wujud dalam ayat yang sama. Fungsi *normalKKMelakukan()*menerima dua input iaitu ayat bagi skema jawapan ke-*N* dan jawapan pelajar ke-*N*. Manakala output bagi fungsi ini merupakan ayat yang menyengaukan kata kerja ‘melakukan’ dengan kata kerja yang lebih signifikan dan dilengkapkan berpandukan kepada petua pembentukan ayat subjek-predikat.

Di peringkat awal proses, kewujudan kesemua token dalam ayat ke-*n* dikenalpasti. Jika didapati wujud kata kerja ‘melakukan’ dalam ayat ke-*n*, maka kata kerja signifikan yang wujud selepas kata kerja ‘melakukan’ tersebut akan dikenalpasti. Jika token yang wujud selepas

kata kerja ‘melakukan’ bersamaan dengan kata ‘proses’, kata tersebut akan dibuang terlebih dahulu kerana tidak memberi impak dalam proses seterusnya. Sebaliknya, setelah kata kerja signifikan yang wujud setelah kata kerja ‘melakukan’ dikenalpasti, kata kerja signifikan pula akan dihuraikan pada imbuhan awalan, kata dasar dan imbuhan akhiran (jika wujud). Jika imbuhan akhiran wujud, maka imbuhan akhir tersebut dibuang daripada kata kerja tersebut.

Setelah itu, proses seterusnya merupakan proses utama dalam fungsi ini iaitu penentuan imbuhan awal yang bakal disengaukan dengan kata dasar kata kerja yang signifikan. Untuk tujuan itu, aksara pertama dan kedua dikenalpasti. Imbuhan awal yang bakal disengaukan adalah imbuhan ‘meN’, yang mana ‘N’ bergantung kepada aksara pertama dan kedua dalam kata dasar kata kerja signifikan tersebut. Malah berdasarkan algoritma tersebut terdapat tiga keadaan berbeza sebelum imbuhan awal berjaya disengaukan dalam kata kerja signifikan iaitu:

- sengauan terus imbuhan awal ke dalam kata kerja signifikan
- huruf pertama dalam kata kerja signifikan perlu digantikan dengan huruf-huruf tertentu terlebih dahulu sebelum sengauan dapat dilakukan
- huruf pertama dalam kata kerja signifikan perlu digugurkan terlebih dahulu sebelum sengauan dapat dilakukan

Akhirnya, kesemua token yang telah dihuraikan di peringkat awal akan digabungkan semula untuk membantuk ayat ternormalisasi berdasarkan kepada petua nahu subjek + (imbuhan awal ‘meN’ + kata kerja signifikan) dan disimpan dalam *normalKKM*.

vi. Penyingkiran perkataan ‘ialah’, ‘adalah’ dan ‘merupakan’ jika terdapat kata kerja yang lebih signifikan wujud dalam sesebuah ayat.

```
Function normalBuangKP():
Input: Ayat bagi jawapan skema jawapan ke-N dan jawapan pelajar ke-N
Output: Ayat dibersihkan daripada kata pemeri
Kenalpasti kewujudan semua token dalam ayat ke-n
If wujud kata pemeri dalam ayat
        If wujud kata kerja signifikan
                Buang kata pemeri dalam ayat
        End If
End if
Simpan ayat yang telah dibersihkan dalam normalBKP
```

Rajah 3.7 Algoritma penyingkiran kata pemeri.

Rajah 3.7 menunjukkan algoritma yang agak ringkas iaitu bertujuan untuk membuang kata pemeri (‘ialah’, ‘adalah’ dan ‘merupakan’) daripada ayat ke-*n*. Fungsi *normalBuangKP()* menerima dua input iaitu ayat bagi skema jawapan ke-*N* dan jawapan pelajar ke-*N*. Manakala output bagi fungsi ini merupakan ayat yang telah dibersihkan daripada kata pemeri.

Di peringkat awal proses, kewujudan kesemua token dalam ayat ke-*n* dikenalpasti. Jika didapati wujud kata pemeri dalam ayat ke-*n*, maka kata kerja signifikan yang wujud selepas kata pemeri tersebut akan dikenalpasti. Jika token yang wujud selepas kata kerja ‘melakukan’ bersamaan dengan kata kerja signifikan, kata pemeri tersebut akan dibuang terlebih dahulu kerana tidak memberi impak dalam proses seterusnya. Akhirnya, ayat yang telah dibersihkan daripada kata pemeri akan disimpan dalam *normalBKP*.

3.3.4 Pembersihan Ayat

Fasa terakhir dalam Fasa Input adalah pembersihan ayat. Ayat-ayat dalam semua dokumen dibersihkan dari segala kata henti dan simbol. Pembersihan ayat perlu dibuat terlebih dahulu kerana beberapa situasi berikut:

i. Nahu bebas konteks digunakan untuk operasi sintaks di mana digunakan untuk menukar token kepada frasa nahu. Token tersebut akan dihuraikan oleh penghurai.