

SRL TOOL: Penentuan Peranan Semantik Berasaskan Heuristik Dalam Pemprosesan Bahasa Tabii

NAZLIA OMAR & SITI SALWA HASBULLAH

ABSTRAK

Penentuan peranan semantik adalah suatu proses untuk menentukan struktur argumen bagi predikat dalam suatu ayat atau teks dengan menggunakan kaedah penentuan semantik. Satu kaedah dalam penentuan peranan semantik ialah dengan menggunakan heuristik. Heuristik merupakan panduan untuk menyelesaikan masalah atau membuat keputusan dengan lebih cepat walaupun kadangkalanya, tidak secara optimum. Kebanyakan alatan atau aplikasi yang sedia ada untuk penentuan peranan semantik adalah terhad dan kurang mengaplikasikan kaedah penggunaan heuristik dalam penentuan tersebut. Kajian ini bertujuan untuk mendapatkan penentuan peranan semantik yang lebih tepat dengan menggunakan kaedah heuristik. Alatan untuk penentuan peranan semantik yang dinamakan SRL TOOL dibangunkan bertujuan untuk meningkatkan ketepatan penentuan peranan semantik menerusi penggunaan heuristik dan rujukan tanda sintaktik. Skop kajian tertumpu kepada penentuan sembilan peranan semantik yang dipilih iaitu Agent, Theme, Goal, Source, Instrument, Beneficiary, Location, Manner dan Time. Pendekatan kaedah penggunaan heuristik adalah untuk menentukan peranan semantik yang wujud dalam suatu ayat Bahasa Inggeris yang diuji. Pengujian dan penilaian telah dilakukan terhadap sembilan peranan semantik yang dikaji dengan menggunakan unit penilaian kejituan, dapatan semula dan F-Measure. Keputusan penilaian menunjukkan dapatan semula yang dicapai ialah 90.0% ,kejituan 81.8% dan F-Measure 85.9% bagi pengujian ke atas set data ujian. SRL TOOL yang dibangunkan ini dapat membantu para penyelidik dalam bidang pemprosesan bahasa tabii. Selain itu juga, pembangunan perisian aplikasi ini turut berguna kepada pakar atau pelajar dalam bidang Linguistik.

Kata kunci: Penentuan peranan semantik, pemprosesan bahasa tabii, heuristik

ABSTRACT

Semantic role labeling is the process of annotating the predicate argument structure in text with semantic labels. One of the technique in semantic role labeling is through the use of heuristics. Heuristics usually serves as a guideline to provide good but not necessarily optimal solutions to difficult problems, easily and quickly. Recent software or tools that assign semantic roles based on heuristics are still minimal. Thus, the aim of this research is to improve the accuracy of semantic role labeling through the use of heuristics. In this research, the semantic role labeling process is performed based on syntactic clues and heuristics. The tool, SRL TOOL is developed using a set of heuristics with the goal to achieve better results in semantic role assigning. The scope of the research covers nine roles in semantic role labeling which include Agent, Theme, Goal, Source, Instrument, Beneficiary, Location, Manner and Time. To further support the contribution for each role, an evaluation of SRL TOOL was carried out. The evaluation measures used are recall, precision and F-Measure. With regard to these evaluations, SRL TOOL has an average of 90.0% recall, 81.8% precision and 85.9% F-Measure for the test dataset. This tool aims to

support the researchers in the Natural Language Processing field. This application tool is also useful for experts or students in Linguistics.

Keywords: Semantic role labeling, natural language processing, heuristic

PENGENALAN

Pemrosesan Bahasa Tabii merupakan kaedah bagaimana komputer digunakan untuk menganalisis dan memahami bahasa manusia. Teknologi pembangunan aplikasi pemrosesan bahasa tabii ini termaksudlah penterjemahan, mendapatkan semula teks, pengkategorian dan rumusan, pengestrakan maklumat dan sistem dialog. Objektif utama pemrosesan bahasa tabii ialah untuk memahami bahasa tabii dalam sebarang domain. Salah satu kaedah untuk memudahkan proses ini ialah dengan menggunakan peranan semantik bagi sesuatu ayat. Hubungan asas penggunaan pemrosesan bahasa tabii dan semantik atau maksud ayat dapat dilihat pada Rajah 1.

Penentuan peranan semantik (*Semantic Role Labeling*) adalah satu cabang dalam bidang pemrosesan bahasa tabii yang berfungsi mencatat struktur hujah predikat dalam sesuatu teks atau ayat dalam suatu dokumen dengan menggunakan penentuan secara semantik (Gildea dan Jurafsky 2000). Terdapat pelbagai jenis kaedah atau pendekatan yang digunakan oleh pengkaji untuk penentuan peranan semantik tersebut. Salah satu kaedah yang boleh digunakan dalam penentuan peranan semantik ini ialah menggunakan heuristik (Omar et. al 2004). Heuristik adalah satu kaedah yang memberikan panduan secara umum bagi sistem atau perisian, untuk mendapatkan maklumat yang berguna dalam membezakan setiap peranan semantik yang diuji (Liu dan Soo 1993).

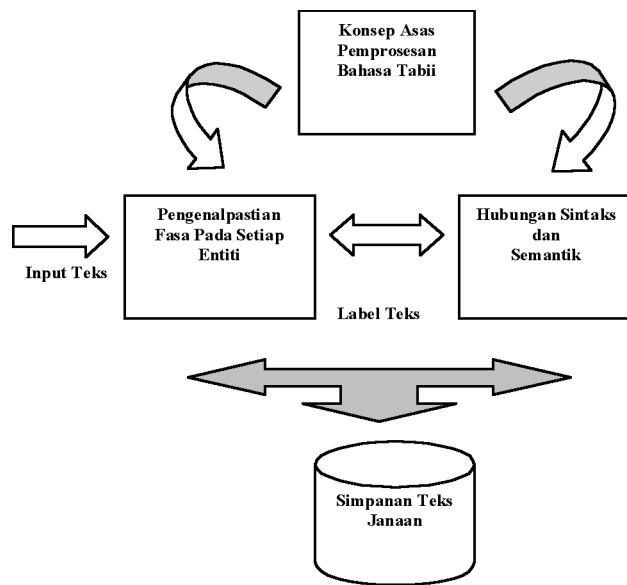
Objektif utama kertas ini adalah untuk mencadangkan heuristik baru dalam menentukan peranan semantik yang wujud dalam suatu ayat Bahasa Inggeris yang diuji. Pembangunan heuristik baru ini adalah berdasarkan satu set data latihan peranan semantik yang telah dikaji daripada aplikasi perisian yang dibangunkan sebelum ini (Fazilah 2006). Pemberat dihasilkan bagi menguji peratusan ketepatan setiap peranan semantik yang diuji. Dengan adanya pembangunan heuristik dan pemberat ini, penentuan peranan semantik dapat dilakukan dengan lebih tepat dan berkesan berbanding dengan anggaran manual yang dilakukan sebelum ini. Pembangunan heuristik ini juga berguna kepada mereka yang menjalankan penyelidikan dalam bidang berkaitan Pemrosesan Bahasa Tabii dan pakar atau pelajar dari bidang Linguistik, yang dapat diaplikasikan dalam persekitaran pendidikan.

LATAR BELAKANG DAN KAJIAN LEPAS

Peranan semantik telah menjadi fokus penyelidikan dalam bidang linguistik sejak beberapa tahun kebelakangan ini. Penggerak utama minat yang mendalam dalam bidang peranan semantik ini terutamanya dalam aplikasi penentuan peranan semantik menggunakan pemrosesan bahasa tabii, ialah keupayaan peranan semantik dianalisis secara automatik (Erk dan Pado 2005). Di antara faedah yang diperolehi daripada aplikasi penentuan peranan semantik ini termasuklah penterjemahan, pemilihan maklumat dan soal jawab sesuatu persoalan bagi sistem.

Peranan semantik ialah sebagai satu bentuk kelas yang agak lama dalam teori bidang linguistik. Peranan semantik ini juga merupakan satu lapisan perhubungan kata kerja (*verb*) utama dalam suatu klausa. Penggunaan set dalam peranan semantik telah dikemukakan oleh ahli linguistik sebagai sebahagian daripada perhubungan sesuatu teori, sebahagian daripada teori grammatikal yang menerangkan tentang sesuatu perhubungan di antara peranan semantik dengan kesedaran peranan semantik itu sendiri.

Dapat diperhatikan bahawa keputusan yang diperolehi di antara teori-teori ini adalah sedikit berbeza dengan set peranan yang digunakan oleh Fillmore(1968), Jackendoff(1990) dan Schank(1972). Di antara contoh peranan semantik yang lazim digunakan adalah seperti dalam Jadual 1.



RAJAH 1. Hubungan asas pemrosesan bahasa tabii dan semantik

Kajian ini didorong dengan penggunaan heuristik sebagai syarat dalam penentuan setiap peranan semantik yang dikaji. Menurut Kamus Dewan Bahasa dan Pustaka (2004), heuristik bermaksud membolehkan atau dapat membantu seseorang mempelajari dan mengetahui sesuatu perkara melalui pengalaman sendiri (bukan kaedah pengajaran atau pembelajaran). Heuristik boleh dibahagikan kepada 4 kaedah yang diperkenalkan oleh Tajularipin (2007) iaitu kaedah heuristik memberi jawapan dengan membuat pilihan jawapan dahulu dan kemudian mencari cara yang sesuai dengan pilihan tersebut. Kaedah heuristik kedua pula, ialah heuristik mudah diingat menerangkan situasi atau peristiwa yang mudah diingat atau baru berlaku akan mempengaruhi dalam membuat keputusan. Kaedah heuristik analogi pula, merujuk cara yang pernah digunakan, yang mempunyai persamaan untuk membuat keputusan dan kaedah keempat ialah heuristik perwakilan yang menggunakan perwakilan dalam membuat rumusan. Kaedah heuristik keempat diperkenalkan oleh Tajularipin (2007) dan kaedah ini adalah bersesuaian dengan kajian yang dilakukan dalam penentuan peranan semantik ini. Perwakilan heuristik akan ditentukan pada setiap perkataan suatu ayat yang diuji dalam menganalisis ayat terpilih pada setiap heuristik yang dibangunkan.

Pada asasnya, perkataan heuristik diterbitkan daripada perkataan Greek iaitu “*heuriskein*” yang bermaksud menemui “*discover*” (Zanakis & Evans 1981). Ianya mencadangkan heuristik boleh digunakan pada sesuatu dengan menyiasat atau menyelidik bagi mendapatkan sesuatu penyelesaian. Selain itu, heuristik juga ditakrifkan sebagai ‘*rules of thumb*’, gerak hati, pertimbangan, ilham, petua yang terbina daripada pengalaman. Panduan untuk menyelesaikan masalah atau membuat keputusan dengan lebih cepat berbanding kaedah rambang (Batra dan Zanakis 1994).

APLIKASI PEMROSESAN BAHASA TABII DALAM PENENTUAN PERANAN SEMANTIK

Terdapat banyak kajian dilakukan yang melibatkan atau menggunakan teknik pemrosesan bahasa tabii untuk menentukan peranan semantik. Gildea dan Jurafsky (2002) menghasilkan satu sistem untuk penakrifan hubungan semantik atau peranan semantik yang diterangkan oleh setiap bahagian (*constituent*) suatu ayat dalam bentuk kerangka semantik (*semantic frame*). Proses penentuan berlaku apabila diberi suatu input bagi ayat dalam bentuk kerangka semantik dan bentuk perkataan, maka sistem akan melabel sebahagian

JADUAL 1. Contoh Peranan Semantik

Peranan Semantik	Maksud	Contoh Ayat
Agent	Penyebab hasrat bagi suatu peristiwa. Orang atau sesuatu yang menyebabkan peristiwa.	<i>He ran down the street.</i> Perkataan <i>He</i> adalah merujuk kepada peranan <i>Agent</i> .
Goal	Destinasi objek bagi peristiwa perpindahan (<i>transfer</i>). Sesuatu yang menunjukkan tindakan adalah langsung.	<i>We put the book on the shelf.</i> Perkataan <i>on the shelf</i> adalah merujuk kepada peranan <i>Goal</i>
Source	Punca asal bagi sesuatu objek bagi peristiwa pemindahan atau penukaran.	<i>Ann borrowed the book from the library.</i> Perkataan <i>from the library</i> adalah merujuk kepada peranan <i>Source</i>
Instrument	Digunakan dalam suatu peristiwa. Merupakan benda bukan hidup yang digunakan oleh <i>agent</i> untuk melaksanakan suatu peristiwa.	<i>Jo cuts hair with a razor.</i> Perkataan <i>a razor</i> adalah merujuk kepada peranan <i>Instrument</i> .
Theme	Merupakan penyebab keberlakuan peristiwa.	<i>The rocket was launched from Central Command.</i> Perkataan <i>The rocket</i> adalah merujuk kepada peranan <i>Theme</i> .
Beneficiary	Yang terlibat mendapat kesan secara langsung daripada peristiwa. Yang mendapat faedah atau keuntungan daripada peristiwa.	<i>Anthony sold the car for a friend.</i> Perkataan <i>for a friend</i> adalah merujuk kepada peranan <i>Beneficiary</i> .
Location	Tempat bagi sesuatu yang dilokasikan.	<i>It rains in Spain.</i> Perkataan <i>Spain</i> adalah merujuk kepada peranan <i>Location</i> .
Time	Memilih atau mengatur masa atau keadaan untuk sesuatu. Tempoh masa keberlakuan peristiwa.	<i>Before 5:25 PM, Erin left.</i> Perkataan <i>5:25 PM</i> adalah merujuk kepada peranan <i>Time</i> .
Manner	Perilaku seseorang terhadap orang lain.	<i>She sang loudly.</i> Perkataan <i>loudly</i> adalah merujuk kepada peranan <i>Manner</i> .

daripada abstrak peranan semantik seperti peranan *Agent*, *Patient* atau juga domain peranan semantik yang lebih spesifik seperti *Speaker*, *Message* dan *Topic*.

Sistem atau aplikasi perisian ini adalah berasaskan pengelasan berstatistik yang diuji secara kasarnya dengan lebih kurang 50,000 ayat yang dilakukan penambahan berasaskan peranan semantik dalam projek semantik “*FrameNet*”. Kemudian, setiap ayat tadi dihurai (*parsed*) ke dalam pokok sintaktik (*syntactic tree*) dan pelbagai ciri leksikal dan sintaktik diambil termasuklah setiap jenis fasa dalam setiap *constituent* iaitu fungsi grammatikal dan juga posisi dalam setiap ayat. Ciri-ciri ini kemudiannya digabungkan dengan pengetahuan predikat dalam suatu kata kerja, kata nama atau kata sifat seperti juga maklumat pada keutamaan

kemungkinan dalam penggabungan peranan semantik yang pelbagai. Mereka juga menggunakan algoritma kelompok leksikal yang pelbagai untuk membuat kesimpulan secara umum bagi mengatasi kemungkinan yang mungkin dalam pengisian peranan. Ayat kemudiannya dihurai (*parsed*), ditambah kepada ciri-ciri (*features*) yang ada dan akhirnya dihantar melalui pengkelasan.

Mereka juga mengatakan bahawa sistem ini mencapai 82% ketepatan dalam pengenalan kepada peranan semantik dalam penghuraian *constituent*. Bahagian yang agak sukar dalam penghurai sesuatu *constituent* dan menerangkan peranan semantik itu, namun kajian ini masih dapat mencapai ketepatan sebanyak 65%.

Kajian ini juga memerlukan mereka membuat perbandingan kegunaan pada setiap ciri yang berbeza dan penggabungan ciri-ciri kaedah dalam tugas peranan semantik. Kajian ini turut mengkaji integrasi pada penentuan peranan yang menghurai sintaktik berstatistik dan cuba membuat satu kesimpulan untuk predikat yang tidak kelihatan pada data latihan. Penggunaan pengkelasan penghurai berstatistik adalah ciri penting dalam pembangunan sistem bagi penentuan peranan semantik kajian yang dilakukan.

Swier dan Stevenson (2004) pula, mencadangkan suatu kaedah tanpa penyeliaan (*unsupervised*) untuk penentuan argumen kata kerja dengan peranan semantik. Satu algoritma yang dipanggil “*bootstrapping*” mengemaskinikan kemungkinan model di mana setiap tugas diasaskan. Kajian ini menjangkakan setiap pendekatan yang menggunakan kata kerja, *slot* dan kata nama sebagai asas untuk sokongan dalam kebarangkalian model. Eksperimen bagi kajian ini mencapai 50-65% pengurangan kadar ralat yang berlaku pada pemberitahuan garis asas (*baseline*). Ia menunjukkan potensi pada pendekatan yang dilakukan untuk mara ke hadapan di mana ianya bergantung kepada data latihan secara besar. Dalam kajian ini, mereka yang membangunkan kaedah baru tanpa penyeliaan penentuan peranan semantik untuk mengelakkan keperluan penentuan teks secara manual yang agak konsisten dan juga membolehkan kegunaan yang besar dan mewakili satu set tulisan. Untuk mencapai peringkat ini, kajian menggunakan pendekatan “*bootstrapping*” di mana pada awalnya hanya tugas peranan sahaja dilakukan dengan jelas berdasarkan kata kerja leksikal. Kemudian, dilakukan pula kebarangkalian model berdasarkan pada penambahan peranan semantik semasa yang mana menggunakan kebarangkalian model ini untuk dikenalpasti peranan yang dianggap mempunyai bukti yang jelas dan ditambahkan pula argumen penentuan yang baru pada setiap set tambahan yang sedia ada.

Kajian ini menerangkan keputusan yang diambil ialah kebarangkalian pada setiap peranan kata kerja, slot sintaktik dan yang berlaku pada kata nama dalam slot tersebut. Kajian ini menggunakan statistik untuk bahagian kelas yang lebih umum pada setiap maklumat, seperti adanya syarat yang dilakukan terhadap kelas semantik pada kata kerja dan juga kepada kata kerja itu sendiri. Pendekatan kajian ini mengambil kemungkinan setiap modelnya menitikberatkan keputusan menggunakan pendekatan pengkelasan secara umum.

Kajian yang dilakukan oleh Liu dan Soo (1993) pula, mencadangkan sebuah kaedah perolehan untuk mendapatkan pengetahuan semantik dengan mempergunakan kata kunci sintaktik daripada ayat latihan (*training sentence*). Kekaburan keputusan pada heuristik digunapakai untuk mengumpul perbezaan lebihan maklumat daripada set data latihan. Untuk memperoleh pengetahuan semantik leksikal, segala argumen dalam peranan semantik pagi setiap ayat dikenalpasti mempunyai bahagian penting sebagai khidmat tindakan saringan bagi mengurangkan jarak hipotesis bagi argumen peranan semantik yang mungkin wujud dalam ayat latihan. Heuristik berasaskan fenomena linguistik yang pelbagai dengan adanya kekangan, diperkenalkan untuk mengatasi permasalahan kekaburan ini. Heuristik ini memberikan bimbingan umum kepada sistem untuk mengumpul maklumat bernilai bagi membezakan peranan semantik.

Teknik yang digunakan dalam kajian ini menggunakan pendekatan syarat aturan dan heuristik dalam penyelesaian kajian mereka. Namun, hanya pendekatan pembangunan secara umum dinyatakan sebagai pengenalan dalam penggunaan heuristik penentuan semantik. Ayat ujian yang diuji dan digunakan oleh pemproses bersintaktik adalah untuk mengekstrak suatu frasa kata yang wujud dalam suatu ayat leksikal. Kajian ini mencadangkan beberapa heuristik yang boleh dilakukan terhadap sistem dalam membantu proses

mendapatkan maklumat dalam mengatasi kekaburan (*ambiguity*) yang wujud dalam pengestrakan suatu frasa ayat. Jadual 2 menunjukkan heuristik yang membezakan setiap peranan *thematic* yang dicadangkan oleh pengkaji dalam mengatasi permasalahan yang timbul.

JADUAL 2. Heuristik dalam membezakan peranan *Thematic*

<i>Volition Heuristic</i> (VH)	Tujuan pembinaan dan tujuan <i>adverbial</i> akan wujud dalam suatu ayat bergandingan dengan argumen pada <i>Agent</i> (Gruber 1976).
<i>Imperative Heuristic</i> (IH)	Heuristik jenis ini hanya diperlukan dalam subjek untuk <i>Agent</i> (Gruber 1976).
<i>Thematic Hierarchy Heuristic</i> (THH)	Menjelaskan mengenai hierarki <i>thematic</i> daripada yang tertinggi kepada yang terbawah iaitu “ <i>Agent > Location, Source, Goal > Theme</i> ”. Frasa ayat pasif berada pada tahap tertinggi daripada subjek terbitan dalam suatu hierarki (contohnya, situasi hierarki <i>thematic</i> seperti yang dicadangkan oleh Jackendoff 1972). Oleh itu, dalam kajian yang dilakukan ini, mereka telah membentuk satu hierarki seperti berikut; <i>Agent > Location, Source, Goal, Instrument, Cause > Theme, Beneficiary, Time, Quantity, Preposition, Manner, Result</i> . Subjek dan objek tidak berada dalam tahap yang dinyatakan ini.
<i>Preposition Heuristic</i> (PH)	Argumen wujud adalah untuk mengemukakan maklumat yang tepat dalam penyelesaian kekaburan suatu <i>thematic</i> .
<i>One-Theme Heuristic</i> (OTH)	Suatu argumen dinyatakan sebagai <i>Theme</i> apabila hanya peranan <i>Theme</i> ini yang mungkin wujud dalam struktur suatu argumen dalam suatu ayat.
<i>Uniqueness Heuristic</i> (UH)	Tidak akan ada 2 jenis argumen wujud dalam suatu peranan semantik (muncul dalam hubungan dan anaphora yang berkait dengan penentuan <i>constituent</i> dengan <i>thematic</i> yang sama.)

Kajian yang dilakukan oleh Punyakanok et al. (2004) pula adalah dengan membangunkan satu sistem dalam penentuan peranan semantik. Sistem yang dibangunkan ini adalah menggabungkan teknik pembelajaran secara mesin (*machine learning*) dengan suatu prosedur berasaskan kesimpulan kepada pengaturcaraan secara *linear* yang menyokong penggunaan hubungan linguistik dan struktur kekangan dalam proses penyelesaian suatu masalah. Sistem yang dibangunkan ini telah diuji daripada persidangan penentuan peranan semantik *CoNNL 2004* dan mendapat keputusan yang agak baik. Data yang diuji di ambil dari korpus *PropBank* yang mana dapat dibezakan prestasi sistem yang dibangunkan dengan sistem lain sediaada.

Dalam kajian ini, mereka menganggap sistem penentuan peranan semantik ini hanya mengambil input maklumat sintaktik separa dan tiada pengetahuan asas berdasarkan leksikal semantik. Secara spesifiknya, sumber input utama adalah diambil daripada penglabelan kelas perkataan (*Part Of Speech Tagger*), iaitu suatu penghurai yang memproses input kepada suatu tahap *chunks* dan klausa dan sebagainya. Namun, mereka tidak menganggap penghurai secara keseluruhan sebagai input dalam proses yang dilakukan.

Inovasi yang dilakukan dalam pendekatan mereka ialah pembangunan kaedah prinsipal iaitu dengan menggabungkan teknik pembelajaran secara mesin dengan linguistik dan struktur kekangan dalam proses penyelesaian suatu masalah. Bagi teknik pembelajaran secara mesin, sistem yang dibangunkan dibahagikan kepada 2 fasa. Fasa pertama ialah suatu set argumen dihasilkan menggunakan 2 jenis pengklasifikasian. Fasa yang pertama digunakan untuk mencari posisi permulaan dan posisi akhir pada setiap jenis argumen yang dicadangkan. Manakala fasa kedua pula, menyatakan satu set argumen daripada fasa pertama diambil kembali menggunakan rekaan pengklasifikasian untuk menentukan jenis argumen yang diberikan.

Ini bermakna, kajian yang dilakukan ini lebih mencadangkan penggunaan pengaturcaraan secara *linear* dalam mempersembahkan kesimpulan akhir yang digunakan dalam penghasilan kajian. Ini dilakukan dengan

menyediakan argumen dalam menyokong penggunaan pengaturcaraan secara *linear* dengan menyatakan peratus capaian pada data ujian yang dilakukan. Kajian juga turut melakukan pengekodan secara *linear* dan penggunaan pengaturcaraan secara *linear* adalah kesimpulan yang dinyatakan sebagai keputusan akhir yang berkesan dalam proses mengatasi sebarang kekangan yang mungkin berlaku. Walaupun secara umumnya, pengaturcaraan secara linear ini sukar untuk dilaksanakan, namun dalam kajian yang dilakukan ini, mereka telah menggunakan pakej pengaturcaraan secara linear (*Xpress-MP, 2003*) dan dapat memproses ayat secara kasarnya sebanyak 20 ayat persaat.

Kajian yang dilakukan oleh Pradhan et al. (2005) menggunakan pendekatan pandangan sintaktik yang berbeza dalam penentuan peranan semantik. Kajian yang dilakukan adalah dengan memperkenalkan garis asas pada teknik terkini yang diperkenalkan dalam penentuan peranan semantik berdasarkan pengklasifikasian secara *Vector Machine*. Dalam kajian yang dilakukan ini, mereka membuktikan peningkatan prestasi kajian ini adalah berasaskan perkara seperti berikut:

- i. Menambahkan ciri-ciri baru termasuklah ciri yang telah diekstrak daripada penghurai yang digunakan.
- ii. Memperkenalkan pemilihan dalam pengukuran suatu penentuan.
- iii. Menggabungkan penghurai yang diperoleh daripada pengujian penghurai semantik yang menggunakan pandangan sintaktik yang berbeza.

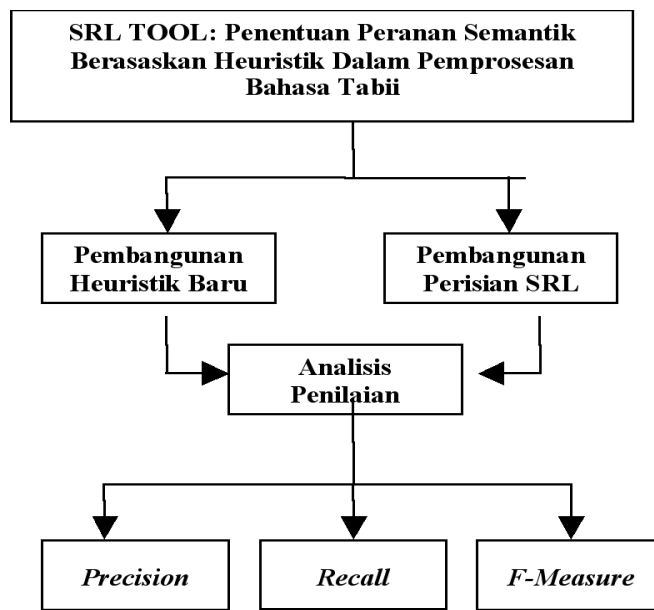
Sistem garis asas yang dikemukakan dalam kajian ini adalah berdasarkan pemilihan teknik terbaik yang diperkenalkan. Setelah itu, kajian merekodkan 2 set eksperimen yang menggunakan teknik meningkatkan prestasi penyelesaian kepada argumen yang dipersembahkan dalam bentuk analisis sintaktik. Pada permulaannya, set eksperimen diuji ciri-ciri barunya, termasuklah ciri-ciri yang diekstrak daripada penghurai yang menyediakan pandangan sintaktik yang berbeza iaitu menggunakan penghurai *Combinatory Category Grammar (CCG)* (Hockenmaier & Steedman 2002).

Pada set eksperimen yang kedua pula, mereka mengkaji pendekatan untuk menakrifkan ciri-ciri subset optimal pada setiap pengklasifikasian argumen dan juga kebarangkalian daripada pengelasan ini. Kajian kemudiannya merekodkan eksperimen yang mengandungi permasalahan yang hilang daripada pandangan analisis secara sintaktik. Jadi, kajian ini mengkaji cara untuk menggabungkan hipotesis yang dijanakan hasil daripada penglabelan kelas perkataan peranan semantik yang diuji melalui pandangan analisis sintaktik yang berbeza iaitu diuji daripada penghurai *Charniak* (Charniak 2000), *Minipar* (Lin 1998), dan yang ketiga penghurai *Shallow Syntactic* (Hacioglu 2004). Jadi, kajian yang dilakukan oleh mereka melihat dari sudut pandangan yang berbeza daripada analisis sintaktik yang dikaji.

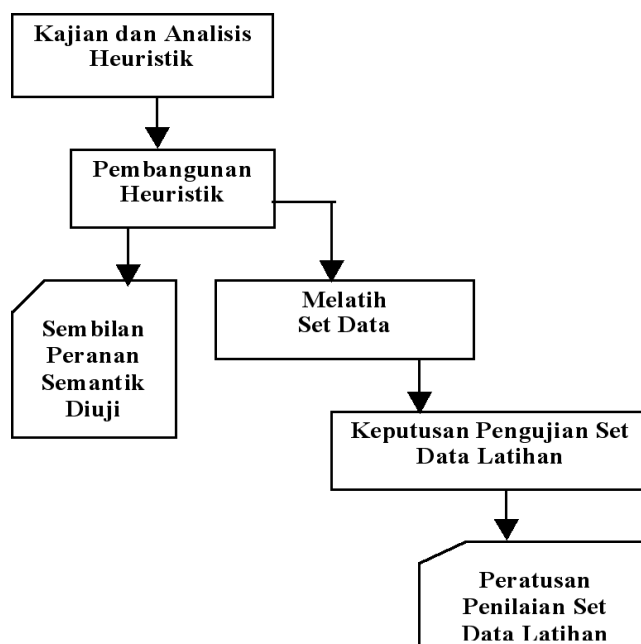
METODOLOGI PEMBANGUNAN KAJIAN

Pembangunan kajian yang dilakukan dibahagikan kepada dua proses utama iaitu pembangunan heuristik baru dan pembangunan perisian aplikasi SRL TOOL. Rajah 2 menunjukkan metodologi kajian yang dilakukan dalam penentuan peranan semantik. Dalam mendapatkan gambaran umum tentang penggunaan heuristik dalam penentuan peranan semantik, kajian yang dilakukan oleh Liu dan Soo (1993) telah mempengaruhi kajian yang dilakukan. Kajian ini mencadangkan satu kaedah perolehan untuk mendapatkan pengetahuan semantik dengan menggunakan kata kunci sintaktik daripada ayat latihan (*training sentence*).

Pembangunan heuristik dalam kajian ini bermula dengan pengumpulan heuristik sedia ada daripada kajian lepas. Satu set data latihan digunakan dalam proses untuk mengenalpasti heuristik baru yang dapat membantu mengenalpasti peranan semantik dengan lebih tepat. Set data latihan ini dikumpul dari ayat yang mengandungi peranan semantik yang jelas dari buku dan kajian lepas. Rajah 3 menunjukkan proses pembangunan heuristik tersebut. Setelah keputusan pengujian set latihan diperoleh, heuristik yang berpotensi dan optima dipilih untuk dibangunkan dalam aplikasi perisian SRL TOOL. Satu set data ujian disediakan untuk menguji sistem ini dalam peringkat pengujian akhir.



RAJAH 2. Metodologi pembangunan kajian



RAJAH 3. Proses pembangunan heuristik

HEURISTIK SEDIA ADA

Heuristik dalam penentuan peranan semantik masih terhad lagi penggunaannya dalam bidang pemprosesan bahasa tabii. Heuristik memberikan gambaran maklumat berasaskan pengalaman. Heuristik boleh digunakan sebagai garis panduan kepada yang tidak berpengalaman untuk melakukan sesuatu yang telahpun digunakan atau dilaksanakan oleh yang berpengalaman. Kajian lepas berkaitan pendekatan heuristik dalam penentuan peranan semantik amat kurang (Fazilah 2006). Namun, terdapat beberapa kajian yang mengutarakan beberapa heuristik secara asas dalam pembangunan kajian mereka. Berikut merupakan heuristik dalam penentuan

peranan semantik bagi peranan *Agent*, *Theme* dan *Time* yang dilakukan oleh Liu dan Soo(1993), Gildea dan Jurafsky (2002), dan Swier dan Stevenson (2004).

Heuristik Penentuan Peranan *Theme*:

Heuristik HTh: Apabila suatu perkataan yang dihuraikan (parsed) sebagai suatu objek dan diikuti oleh frasa kata kerja, maka peranan Theme wujud (Gildea dan Jurafsky, 2002).

Sebagai contoh ayat yang menerangkan heuristik ini adalah seperti di bawah:

“The ruling changed because of the protests”

Dalam heuristik ini, Gildea dan Jurafsky (2002) menggunakan suatu penghurai (*parser*) dalam menentukan objek dan subjek perkataan dalam setiap ayat yang diuji. Jadi, dalam ayat tersebut, perkataan *ruling* adalah dikatakan sebagai objek dan diikuti oleh *changed* sebagai frasa kata kerja. Maka peranan *Theme* wujud.

Heuristik Penentuan Peranan *Time*:

Heuristik HTi: Jika ayat mempunyai set T di mana $T = \{ at, of, in, before, after, by, on, during \text{ dan } about \}$ yang wujud sebelum set R di mana $R = \{ “.”, “:” \}$ maka peranan Time wujud pada ayat dalam set R. (Liu dan Soo 1993).

Sebagai contoh ayat yang menerangkan heuristik ini adalah seperti di bawah:

“At 2:30 Erin left”

Berdasarkan heuristik yang dinyatakan di atas, At adalah perkaaan sebelumnya yang merujuk kepada masa 2.30 dan terdapat simbol “:” yang menyatakan peranan Time wujud.

Heuristik Penentuan Peranan *Agent*:

Heuristik HAg: Apabila suatu perkataan yang dihuraikan(parsed) sebagai suatu objek dalam suatu ayat dan diikuti oleh kata kerja yang telah dikelaskan dalam bentuk kerangka (frame) dan VerbNet, maka peranan Agent wujud (Swier dan Stevenson 2004).

Sebagai contoh ayat yang menerangkan heuristik ini adalah seperti di bawah:

“Yuka whispered to Dar”

Berdasarkan ayat di atas, *Yuka* merujuk kepada objek dan diikuti oleh frasa kata kerja yang dikelaskan dalam bentuk *frame* dan *VerbNet* yang mana kajian mereka perlu mengelaskan perkataan *whispered* dalam bentuk *frame* dan *VerbNet* terlebih dahulu, maka peranan *Agent* dikatakan wujud.

CADANGAN HEURISTIK BARU

Berdasarkan keperluan untuk mempertingkatkan ketepatan keputusan maka satu set tambahan heuristik baru diperkenalkan. Heuristik ini boleh digambarkan sebagai heuristik sintaktik tambahan yang mana boleh

membantu dalam mempertingkatkan lagi keputusan dalam penentuan setiap peranan semantik daripada spesifikasi keperluan bahasa tabii iaitu Bahasa Inggeris. Setiap kesembilan peranan semantik yang dikaji akan dibangunkan heuristik bagi setiap satunya.

Heuristik Baru Penentuan Peranan *Agent*:

Heuristik HA1: Jika suatu ayat mempunyai frasa kata nama atau penentu(Determiner) dan kemudiannya diikuti oleh frasa kata kerja bukan “BE”(VerbBe), maka peranan Agent wujud.

Keutamaan: 50

Sebagai contoh ayat untuk heuristik ini adalah seperti berikut:

“Jack ran down the street”

Berdasarkan contoh ayat bagi peranan *Agent* tersebut, perkataan *Jack* adalah suatu frasa kata nama dan perkataan yang seterusnya iaitu *ran* adalah frasa kata kerja bukan “BE”(Verb BE) di mana frasa kata kerja “BE” adalah seperti *was, is, were* dan *are*. Keutamaan pula diberikan sebanyak 50 adalah untuk menunjukkan pemilihan heuristik yang dilakukan oleh SRL TOOL dalam menentukan peranan semantiknya. Semakin tinggi nilai keutamaan yang diberikan maka semakin tinggi keutamaan heuristik tersebut digunakan untuk menentukan peranan semantik yang diuji. Jadi, ayat yang diuji seperti contoh yang diberikan akan ditentukan sebagai peranan *Agent* berdasarkan heuristik yang dibangunkan.

Heuristik Baru Penentuan Peranan *Beneficiary*:

Heuristik HB: Apabila terdapat perkataan(string) “for” dalam ayat dan kemudiannya diikuti oleh frasa penentu(determiner) dan seterusnya frasa kata nama(noun phrase) maka akan diambil sebagai peranan Beneficiary.

Keutamaan: 50

Sebagai contoh ayat untuk heuristik ini adalah seperti berikut:

“Anthony sold the car for a friend”

Seperti yang ditunjukkan pada ayat di atas, perkataan *for* wujud dalam ayat yang diuji, maka syarat seterusnya mengambil peranan dalam penentuan peranan *Beneficiary* ini. Pemilihan heuristik ini adalah sebanyak 50 juga bertindak sebagai heuristik utama bagi penentuan peranan *Beneficiary* ini.

Heuristik Baru Penentuan Peranan *Location*:

Heuristik HL: Apabila terdapat suatu lokasi atau tempat yang diterbitkan daripada komponen Gazetteer (Location) dalam SRL TOOL maka peranan Location wujud.

Keutamaan: 50

Sebagai contoh ayat untuk heuristik ini adalah seperti berikut:

“Kangaroos inhabit Australia”

Seperti yang ditunjukkan pada ayat di atas, lokasi yang diterbitkan dalam komponen *Gazetteer (Location)* yang boleh dilarikan dalam SRL TOOL adalah merujuk kepada perkataan *Australia* dan diberi keutamaan iaitu sebanyak 50, jadi ditentukan di sini bahawa peranan *Location* wujud dalam ayat tersebut.

Heuristik Baru Penentuan Peranan *Time* :

Heuristik HTi: Apabila terdapat suatu perkataan yang merujuk kepada masa diterbitkan daripada komponen Gazetteer(Date) dalam SRL TOOL maka peranan Time wujud.

Keutamaan: 50

Sebagai contoh ayat untuk heuristik ini adalah seperti berikut:

“Last Tuesday was really fun”

Seperti yang ditunjukkan pada ayat di atas, perkataan yang merujuk kepada masa adalah *Last Tuesday* yang mana diterbitkan daripada komponen *Gazetteer(Date)* yang memberikan keutamaan utama iaitu sebanyak 50 dalam penentuan bagi peranan *Time* ini.

Heuristik Baru Penentuan Peranan *Manner*:

Heuristik HM: Jika suatu ayat mempunyai frasa kata keterangan(Adverb) wujud dalam ayat yang diuji, maka peranan Manner wujud.

Keutamaan: 50

Sebagai contoh ayat untuk heuristik ini adalah seperti berikut:

“They walk slowly”

Berdasarkan pada contoh ayat yang diberikan, perkataan *slowly* merupakan kata keterangan dan ayat tersebut memberikan keutamaan utama dalam peranan *Manner* ini iaitu sebanyak 50.

Heuristik Baru Penentuan Peranan *Goal*:

Heuristik HG: Jika sesuatu label TO terdapat dalam ayat dan kemudiannya diikuti oleh frasa penentu(determiner) serta frasa kata nama(noun phrase), maka peranan Goal wujud.

Keutamaan: 50

Sebagai contoh ayat untuk heuristik ini adalah seperti berikut:

“Horacio swam to the raft”

Berdasarkan pada contoh ayat di atas, perkataan *to* merujuk kepada label TO dan diikuti oleh frasa penentu iaitu perkataan *the* dan disertai pula oleh frasa kata nama iaitu *raft*. Ini adalah heuristik utama yang digunakan dalam menentukan peranan *Goal* dan diberikan nilai keutamaan sebanyak 50.

Heuristik Baru Penentuan Peranan *Instrument*:

Heuristik HI: Jika terdapat perkataan(string) “with” dan diikuti oleh frasa penentu(determiner) serta diikuti pula oleh frasa kata nama(noun phrase) maka peranan Instrument wujud.

Keutamaan: 50

Sebagai contoh ayat untuk heuristik ini adalah seperti berikut:

“Saggion cuts his hair with a razor”

Berdasarkan pada contoh ayat di atas, perkataan “with” wujud dalam ayat dan diikuti oleh perkataan *a* sebagai frasa penentu dan perkataan *razor* sebagai frasa kata nama. Heuristik utama ini memberikan nilai keutamaan sebanyak 50 dalam penentuan bagi peranan *Instrument* ini.

Heuristik Baru Penentuan Peranan *Source*:

Heuristik HS: Jika terdapat perkataan(string) “from” dalam suatu ayat dan diikuti oleh frasa kata nama(noun phrase) maka peranan Source wujud.

Keutamaan: 50

Sebagai contoh ayat untuk heuristik ini adalah seperti berikut:

“Nick grabbed the chocolate from Tom”

Berdasarkan heuristik pertama ini yang menyatakan perkataan *from* dan diikuti oleh frasa kata nama iaitu yang merujuk kepada perkataan *Tom*. Heuristik utama ini memberikan nilai keutamaan sebanyak 50 dalam penentuan bagi peranan *Source* ini.

Heuristik Baru Penentuan Peranan *Theme*:

Heuristik Th: Jika suatu ayat mempunya frasa kata nama(noun phrase), atau frasa kata nama ganti diri(Personal Pronoun) dan diikuti oleh frasa kata kerja(verb phrase), tetapi tidak diterbitkan oleh komponen Gazetteer(Person) maka peranan Theme wujud.

Sebagai contoh ayat dalam heuristik ini adalah seperti berikut:

“The caravan continued on towards the distance oasis”

Heuristik bagi peranan *Theme* yang mana heuristik yang terakhir peranan semantik yang dikaji hanya membangunkan satu heuristik sahaja berdasarkan fakta yang telah diselidiki. *The caravan* adalah merujuk kepada peranan *Theme* yang diwakilkan oleh frasa kata nama dan diikuti oleh frasa kata kerja iaitu perkataan *continued*. Namun syarat yang diberikan dalam pembangunan heuristik adalah tidak menggunakan terbitan *Gazetteer (Person)* dalam penentuan peranan *Theme* ini. Oleh kerana peranan *Theme* ini hanya diwakili oleh satu heuristik, maka aturcara tidak meletakkan keutamaan dalam membezakan penentuan peranan semantik ini seperti yang telah ditunjukkan dalam heuristik di atas.

Jadual 3 menunjukkan pembangunan heuristik baru dan heuristik yang sedia ada dalam kajian yang dilakukan ini. Dalam jadual tersebut, dapat diperhatikan bahawa kajian yang dilakukan terhadap pembangunan

heuristik sebelum ini amatlah kurang menggunakan dalam melakukan penentuan pada setiap peranan semantik ini. Berdasarkan jadual tersebut, hanya peranan *Agent*, *Theme* dan *Time* sahaja mempunyai heuristik dalam penentuan peranan semantik tersebut. Namun, kajian ini mempertaruhkan sebilangan besar heuristik pada setiap peranan semantik yang dikaji. Sebanyak 21 set heuristik telah dihasilkan sepanjang kajian yang dilakukan terhadap kesemua sembilan peranan semantik yang diuji.

JADUAL 3. Pembangunan Heuristik Baru dan Sedia Ada

Peranan Semantik	Heuristik Sedia Ada	Heuristik Baru
<i>AGENT</i>	1	4
<i>GOAL</i>	-	3
<i>SOURCE</i>	-	3
<i>INSTRUMENT</i>	-	2
<i>THEME</i>	1	1
<i>BENEFICIARY</i>	-	2
<i>LOCATION</i>	-	2
<i>TIME</i>	1	2
<i>MANNER</i>	-	2

SENI BINA APLIKASI PERISIAN SRL TOOL

Satu seni bina sistem adalah reka bentuk konseptual atau kumpulan hubungan antara sesuatu bahagian dalam pembangunan sesebuah sistem atau perisian. Dalam menentukan peranan semantik data ujian, kaedah taabiran (*inference*) rantaian ke hadapan digunakan. Ini adalah kerana kaedah ini bermula dengan proses padanan satu set fakta (syarat) yang diketahui dan berterusan sehingga matlamat (peranan) ditemui. Dalam menentukan peranan semantik ayat-ayat atau data-data ujian menggunakan SRL TOOL yang dibangunkan ini, terdapat beberapa peringkat yang terlibat. Rajah 4 menunjukkan seni bina sistem SRL TOOL yang menggambarkan fungsi keseluruhan SRL TOOL yang dilaksanakan menggunakan aplikasi GATE, iaitu satu seni bina umum teknologi kejuruteraan berasaskan teks dalam pemprosesan bahasa tabii.

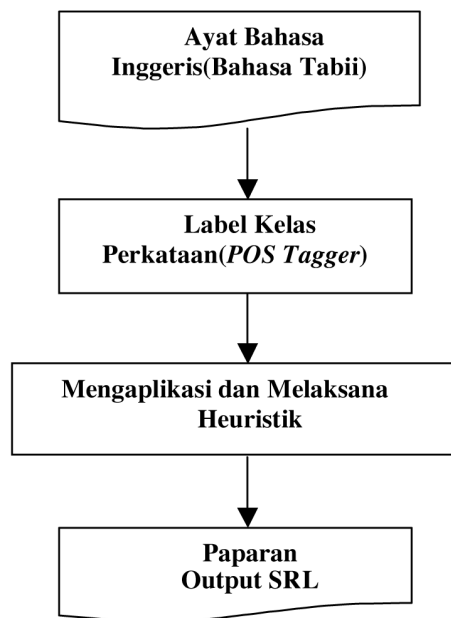
Terdapat beberapa langkah yang perlu dilakukan atau dipenuhi bagi menghasilkan atau memaparkan hasil akhir atau output yang lebih tepat bagi menentukan setiap peranan semantik. Antara langkah-langkah tersebut adalah seperti berikut:

- i. Mendapatkan fail input iaitu ayat diuji dalam Bahasa Inggeris (Bahasa Tabii).
- ii. Penglabelan kelas perkataan (POS *Tagger*)
- iii. Mengaplikasi dan melaksanakan heuristik
- iv. Penentuan peranan semantik dan paparan output

SENI BINA GATE

General Architecture Text Engineering (GATE) versi 4.0 adalah satu seni bina, persekitaran pembangunan dan juga satu rangka kerja dalam pembangunan sistem yang digunakan dalam memproses bahasa tabii dan telah dibangunkan sejak dari tahun 1996 (Cunningham 2002).

GATE versi 4.0 ini telah dibangunkan oleh penyelidik daripada University of Sheffield dan digunakan pada kebanyakan projek penyelidikan (Maynard 2000), termasuklah bidang carian maklumat berasaskan pelbagai bahasa dan perantaraan serta turut digunakan dalam bidang pemprosesan bahasa tabii. Seni bina GATE ini adalah antara satu sistem yang banyak digunakan dengan meluas kerana ianya adalah satu seni



RAJAH 4. Senibina bagi SRL TOOL

bina infrastruktur untuk pembangunan perisian berasaskan pemprosesan bahasa tabii. GATE membantu pembangun, pakar dan penyelidik dalam bidang pemprosesan bahasa dalam tiga cara seperti berikut:

- i. Menentukan satu seni bina, atau satu struktur organisasi untuk perisian pemprosesan bahasa.
- ii. Menyediakan satu rangka kerja, atau *class library* yang mengimplementasikan satu seni bina yang boleh mempertingkatkan keupayaan pemprosesan bahasa dalam pelbagai aplikasi.
- iii. Turut menyediakan satu persekitaran pembangunan yang dibina di atas suatu rangka kerja yang memberi kemudahan kepada alatan grafik (*graphic tools*) dalam pembangunan komponen yang dilakukan.

Penggunaan seni bina GATE (Cunningham 2002) turut berfungsi dengan mendapatkan keputusan saintifik yang relevan pada hasil akhir kajian yang dilakukan dalam bidang seperti pengkomputeran linguistik, pemprosesan bahasa tabii dan kecerdasan buatan secara umumnya. Terdapat beberapa komponen dalam senibina GATE seperti ANNIE (A Nearly-New Information Extraction System) yang menyokong pembangunan aplikasi perisian SRL TOOL yang bertindak sebagai platform yang baik dalam penentuan peranan semantik ini.

A NEARLY-NEW INFORMATION EXTRACTION SYSTEM (ANNIE)

Seni bina GATE versi 4.0 menyokong dan diintegrasikan dengan suatu sistem carian maklumat (*Information Extraction*) yang dipanggil ANNIE (A Nearly-New Information Extraction System). ANNIE juga dikenali sebagai suatu komponen atau satu modul dalam senibina GATE versi 4.0 yang bergantung pada alkhawarizmi yang telah dinyatakan dan bahasa yang digunakan dalam pembangunannya adalah JAPE (*Java Annotations Pattern Engine*). Dengan kata lain, komponen ANNIE berguna dalam aplikasi yang berbeza, dalam melakukan kajian pada suatu teks yang dikaji untuk setiap tujuan yang berlainan (Maynard 2000). Bagi mendapatkan suatu pengujian terhadap teks atau dokumen dalam komponen ANNIE ini, terdapat beberapa sub modul atau komponen yang menjalankan fungsi ANNIE dengan sempurna mengikut keperluan teks yang ingin diuji dan ditentukan ini. ANNIE mengandungi beberapa komponen yang digunakan oleh aplikasi perisian SRL TOOL sebagai satu platform untuk menguji peranan semantik seperti *Tokeniser*, *Gazetteer*, *Sentence Splitter*, *Part Of Speech Tagger* (Hepple 2000) dan *SUPPLE Parser*.

PENILAIAN

Kebanyakan penyelidikan yang dijalankan dalam bidang pemprosesan bahasa tabii dan juga capaian maklumat dalam era sedekad yang lalu telah dihubungkan dengan kaedah asal yang diperkenalkan oleh MUC (1995). Penilaian yang dikemukakan menggunakan unit penilaian kejitian (*precision*), dapatan semula (*recall*) dan F-Measure turut digunakan dalam menentukan penilaian yang mempunyai sedikit variasi baru. Unit penilaian ini merupakan satu unit penilaian yang sudah lama digunakan dalam bidang pemprosesan bahasa tabii dan juga capaian maklumat (Van Rijsbergen 1979).

Unit penilaian bagi kejitian berfungsi untuk mengukur jumlah ketepatan set data ujian yang dikenalpasti dalam bentuk peratusan daripada jumlah bilangan set data ujian yang diuji. Dengan kata lain, unit penilaian ini mengukur berapa banyak ayat ujian yang diuji oleh SRL TOOL yang dikenalpasti adalah betul tanpa menghiraukan sama ada sistem gagal untuk mengenal pasti ayat ujian diuji dengan betul. Semakin tinggi unit penilaian kejitian yang dilakukan, semakin tinggi ketepatan yang dilakukan oleh sistem dalam menentukan ketepatan ayat-ayat yang diuji dalam SRL TOOL.

Kadar ralat (*error rate*) pula, adalah ketepatan sonsangan (*inverse of precision*), iaitu yang mengukur bilangan ayat-ayat diuji yang dikenal pasti dan dijadikan sebagai peratusan kepada ayat-ayat yang telah diuji. Ia kadangkalanya, digunakan sebagai satu kaedah lain bagi unit penilaian kejitian. Selain itu, unit penilaian dapatan semula pula, mengukur bilangan set data ujian diuji yang tepat untuk dijadikan sebagai peratusan bagi kesemua set data ujian diuji dengan tepat. Dengan kata lain, unit penilaian ini mengukur berapa banyak set data ujian yang dikenal pasti tanpa menghiraukan berapa banyak pengenalpastian yang salah telah dilakukan. Jadi, semakin tinggi kadar penilaian bagi dapatan semula, maka semakin tinggi ketepatan yang dilakukan oleh SRL TOOL dengan tidak menghilangkan set data ujian yang telah diuji dengan tepat.

Secara jelasnya, terdapat beberapa cara yang dilakukan di antara kejitian dan dapatan semula bagi memudahkan SRL TOOL untuk mencapai peratusan sehingga 100% kejitian. Ini boleh dilakukan dengan mencapai 100% tanpa melakukan sebarang kesilapan terhadap data yang telah ditentukan atau mencapai 100% bagi dapatan semula dengan tanpa melakukan sebarang kesilapan terhadap data yang telah ditentukan. Manakala menurut Van Rijsbergen (1979) unit penilaian F-Measure pula adalah salah satu unit penilaian yang biasa digunakan oleh unit penilaian kejitian dan dapatan semula yang dirumuskan sebagai pemberat atau purata kepada jumlah yang ditentukan oleh kedua-dua unit penilaian tersebut. Metrik atau unit penilaian yang lazim digunakan dalam menentukan penilaian dalam pemprosesan bahasa tabii ini yang juga penentuan penilaian bagi SRL TOOL adalah seperti diterangkan di bawah :

KEJITUAN

Dalam ukuran metrik penilaian, unit penilaian utama yang digunakan adalah kejitian, dapatan semula dan F-Measure. Unit penilaian bagi kejitian adalah dengan mengukur jumlah bilangan betul dan dibahagikan dengan jumlah hasil bilangan betul dan bilangan salah. Formula yang digunakan dalam kajian ini bagi unit penilaian kejitian adalah seperti di bawah:

$$Kejitian = \frac{NBetul}{NBetul + NSalah}$$

DAPATAN SEMULA

Kajian yang dilakukan dalam menilai prestasi SRL TOOL perlu melakukan penilaian terhadap unit penilaian dapatan semula yang bergandingan dengan kejitian dalam menghasilkan pengukuran prestasi sistem. Unit penilaian bagi dapatan semula diberikan sebagaimana formula di bawah iaitu hasil dapatan semula ditentukan dengan jumlah bilangan betul yang dibahagikan dengan jumlah sebenar ayat yang diuji.

$$Dapatan\ semula = \frac{NBetul}{NSebenar}$$

F-MEASURE

Dalam mendapatkan rumusan atau peratusan akhir metrik penilaian kajian, formula F-Measure dilakukan bagi mencapai hasil akhir penilaian sistem. Nilai unit penilaian bagi F-Measure diukur sebagaimana yang dinyatakan seperti formula di bawah. Simbol β dalam formula di bawah digambarkan sebagai pemberat kejituan dan dapatan semula, di mana jika β adalah bersamaan dengan 1, maka kedua-dua pemberat kejituan dan dapatan semula akan bertambah.

$$F - Measure = \frac{(\beta^2 + 1) \text{Kejituan} * \text{Dapatan}}{(b^2 \text{Dapatan}) + \text{Kejituan}}$$

KEPUTUSAN PENILAIAN

Setelah melakukan pengujian dan penilaian terhadap SRL TOOL, keputusan penilaian yang diperoleh memaparkan penilaian akhir bagi setiap peranan semantik yang ditentukan. Keputusan penilaian yang dilakukan ke atas set data ujian ini, menggunakan sebanyak 75 ayat ujian baru meliputi kesemua peranan semantik yang diuji. Setiap peranan semantik ditentukan melalui setiap ayat yang diuji peranan semantiknya dan memberi keputusan akhir berdasarkan heuristik yang telah dibangunkan. Keputusan penilaian ini penting bagi mengukur keberkesanan dan ketepatan SRL TOOL dalam menentukan setiap peranan semantik yang telah dibangunkan.

Jadual 4 menunjukkan bilangan set data ujian atau ayat yang dimasukkan bagi menguji setiap peranan semantik untuk menentukan ketepatannya. Sebanyak 75 set data ujian baru diuji bagi menilai ketepatan aplikasi SRL TOOL berdasarkan heuristik yang telah dibangunkan. Kesemua set data ujian yang diuji adalah set data ujian yang baru dan dipilih berdasarkan peranan semantik yang tersendiri. Ini adalah penting kerana aplikasi SRL TOOL hanya dapat menentukan peranan semantik hasil daripada set data ujian yang mempunyai hubungan semantik tersendiri. Ianya juga sama sekali tidak boleh menggunakan sebarang ayat yang tidak mempunyai erti atau tidak berkaitan dengan semantik dan bahasa tabii yang telah ditekankan dalam pembelajaran dalam bidang Linguistik ini. Setelah pemilihan set data ujian atau ayat dilakukan, maka setiap data ujian ini seterusnya diuji peranan semantik masing-masing dan ditentukana ketepatannya melalui aplikasi SRL TOOL ini. Jadi, berdasarkan Jadual 4, didapati sebanyak 69 ayat atau set data ujian yang baru diuji adalah tepat merujuk kepada setiap peranan masing-masing hasil daripada 75 jumlah bilangan kesemua data ujian yang diuji.

Jadual 5 pula, menunjukkan rumusan keputusan akhir dalam pengujian set data bagi setiap peranan semantik yang diuji. Hasil keputusan akhir penilaian ini dikemukakan dalam bentuk kejituan, dapatan semula dan F-Measure. Penentuan peranan semantik ke atas aplikasi SRL TOOL memerlukan keputusan penilaian bagi mendapatkan hasil yang lebih baik. Keputusan penilaian daripada Jadual 5 digunakan dalam formula yang telah dinyatakan yang mana memaparkan keputusan yang diperoleh bagi setiap peranan semantik yang diuji berdasarkan kepada set data ujian. Penentuan setiap peranan semantik memerlukan keputusan penilaian bagi mengukur ketepatan setiap peranan yang diuji. Unit penilaian yang digunakan untuk mengukur ketepatan peranan semantik ini adalah kejituan, dapatan semula dan F-Measure.

Berdasarkan keputusan yang dipaparkan dalam Jadual 5, didapati bahawa aplikasi SRL TOOL menghasilkan nilai dapatan semula yang tinggi iaitu sebanyak 90.0%. Pembangunan SRL TOOL berjaya menentukan kesemua peranan semantik bagi mendapatkan penentuan yang lebih tepat. Ini juga menunjukkan bahawa heuristik yang digunakan adalah menepati ciri yang diperlukan dalam menentukan setiap peranan semantik yang diuji. Selain itu, dapat dilihat berdasarkan kepada nilai dapatan semula yang tinggi tersebut,

JADUAL 4. Keputusan Ketepatan Bilangan Set Data Ujian

Peranan Semantik	Bilangan Data Ujian	Bilangan Data Ujian (Tepat)
Agent	9	8
Goal	6	5
Source	13	11
Instrument	5	5
Theme	8	7
Beneficiary	5	4
Location	7	7
Time	13	13
Manner	9	9
Purata	75	69

JADUAL 5. Keputusan Penilaian SRL TOOL Bagi Setiap Peranan Semantik

Peranan Semantik	Kejituan	Dapatan Semula	F-Measure
Agent	89.0%	89.0%	89.0%
Goal	83.0%	83.0%	83.0%
Source	90.0%	75.0%	82.5%
Instrument	56.0%	100%	78.0%
Theme	100%	80.0%	90.0%
Beneficiary	63.0%	83.0%	73.0%
Location	88.0%	100%	94.0%
Time	10%	100%	100%
Manner	67.0%	100%	83.5%
Purata	81.8%	90.0%	85.9%

SRL TOOL yang berasaskan heuristik ini mampu menjana penentuan peranan semantik yang tepat dalam pembelajaran dalam bidang linguistik dan juga pemprosesan bahasa tabii ini.

Nilai yang diperoleh bagi kejituan pula adalah sebanyak 81.8%. Ini menepati setiap set data ujian yang telah diuji dan mempunyai ketepatan ujian yang juga tinggi. Ini adalah kerana, kejituan merujuk kepada ketepatan penentuan peranan semantik pada set data ujian. Namun terdapat beberapa ayat yang diuji tidak relevan dengan heuristik yang dinyatakan. Namun begitu, walaupun kejituan sedikit berkurangan berbanding dapatan semula, namun peratusan keseluruhan masih berada pada kedudukan yang tinggi iaitu sebanyak 85.9% yang merujuk kepada F-Measure.

Pengukuran penilaian pada unit F-Measure ini adalah hasil akhir bagi kedua-dua unit penilaian dapatan semula dan kejituan. F-Measure merujuk kepada jumlah atau purata keseluruhan yang telah dijanakan oleh kejituan dan dapatan semula. Ini dapat disimpulkan bahawa sebanyak 85.9% telah dijanakan oleh F-Measure daripada penentuan peranan semantik aplikasi SRL TOOL, dan peratusan yang didapati ini menunjukkan bahawa heuristik yang dibangunkan dapat menentukan peranan semantik yang baik dan tepat. Melalui pembangunan heuristik yang pelbagai pada setiap peranan semantik, maka aplikasi SRL TOOL dapat

melakukan pemilihan heuristik sebelum menentukan peranan semantik dan cara ini adalah tepat untuk mendapatkan hasil akhir yang lebih baik.

PERBANDINGAN HASIL KAJIAN

Setelah segala kajian dilakukan dalam mencapai matlamat penentuan peranan semantik yang baik, maka keputusan penilaian yang diperoleh daripada proses pengujian SRL TOOL dibandingkan dengan kajian yang terdahulu. Perbandingan ini perlu dalam menguji tahap keberkesanan heuristik yang dibangunkan agar mempunyai prestasi yang setaraf dan mampu menyaingi aplikasi kajian lepas. Kajian terdahulu yang dilakukan perbandingan ialah kajian yang dilakukan oleh Liu dan Soo (1993) kerana ianya merupakan satu-satunya kajian yang menggunakan heuristik dalam penentuan peranan semantik. Namun perbandingan tidak dapat dilakukan secara kuantitatif kerana set data ujian yang digunakan adalah berlainan.

Liu dan Soo (1993) dalam kajian yang dibangunkan mencadangkan satu kaedah perolehan untuk mendapatkan pengetahuan semantik dalam penentuan peranan semantik. Kajian mereka lebih menumpukan penggunaan kata kunci sintaktik daripada ayat latihan. Penggunaan heuristik dalam kajian yang dilakukan adalah untuk mengumpul perbezaan maklumat daripada set latihan. Namun, kajian mereka dikatakan hanya memfokus kepada pembangunan heuristik secara umum dan heuristik yang dibangunkan oleh mereka tidak didasarkan kepada setiap peranan semantik seperti yang dilakukan oleh SRL TOOL. Mereka hanya mensasarkan heuristik secara pengenalan untuk pembangunan kajian yang lanjut dilakukan.

Selain itu juga, kajian yang dilakukan oleh Pradhan et. al (2005) tidak menggunakan pendekatan heuristik dalam penentuan peranan semantik. Teknik garis asas yang digunakan hanya untuk mengklasifikasikan setiap ayat yang diuji untuk dipadankan kepada ciri-ciri subset secara optimal namun tidak mengekstrakkan kajian mereka kepada heuristik. Begitu juga, kajian yang dilakukan oleh Gildea dan Jurafsky (2002), menghasilkan sistem penentuan peranan semantik berdasarkan klasifikasi berstatistik. Dalam merungkaikan masalah kekaburan dalam suatu ayat yang diuji, kajian yang dilakukan ini menggunakan alkhwarizmi kelompok leksikal yang pelbagai untuk membuat kesimpulan secara umum dalam mengatasi kemungkinan atau kekaburan dalam masalah penentuan sesuatu peranan semantik. Jadi, tidak dinyatakan di sini penggunaan heuristik dalam mengatasi kekaburan dalam membuat kesimpulan penentuan peranan semantik.

Berdasarkan pendekatan heuristik yang kurang digunakan dalam kajian terdahulu, maka langkah pembangunan heuristik bagi aplikasi perisian SRL TOOL adalah tepat dalam mendapatkan penentuan yang lebih baik pada setiap peranan semantik yang diuji. Setiap peranan semantik yang diuji dilakukan analisis dan dibangunkan heuristik bagi menguji prestasi aplikasi perisian SRL TOOL dalam menjalankan fungsinya.

KESIMPULAN

Secara keseluruhannya, kajian terhadap peranan semantik yang dilaksanakan ini berjaya mencapai objektifnya. Proses penentuan peranan semantik berjaya dilakukan berasaskan penggunaan heuristik bagi setiap penentuan semantik tersebut. Setelah melakukan beberapa ujian terhadap pelbagai set data atau ayat yang dikenalpasti peranan semantiknya, aplikasi ini berjaya menghasilkan output yang baik dari segi aspek kebolehpercayaannya. Keputusan penilaian yang diperoleh semasa proses pengujian SRL TOOL telah menyokong objektif utama kajian bahawa pembinaan heuristik dapat menyumbang kepada pembinaan SRL TOOL yang dapat menentukan kesemua peranan semantik dengan lebih baik. Satu set heuristik telah dibangunkan untuk menentukan peranan semantik dengan memerhatikan ciri-ciri yang sama bagi sesuatu peranan yang terdapat pada kebanyakan ayat asal yang diuji (ayat tersebut telah diketahui atau dikenalpasti peranan semantiknya). Pembangunan aplikasi perisian SRL TOOL ini menjadi tunjang utama yang menggunakan kaedah heuristik dalam penentuan peranan semantik. Kajian ini turut menjadi pemangkin untuk kegunaan masa hadapan seandainya cadangan perluasan projek yang telah dikemukakan terlaksana.

RUJUKAN

- Batra, D. & Zanakis, S. H. 1994. A Conceptual Database Design Methodology Based on Rules and Heuristics. *European Journal of Information Systems* 3(3): 228-239.
- Charniak, E. 2000. A Maximum-Entropy-Inspired Parser. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, hlm. 132–139.
- Cunningham, H. 2002. GATE, A General Architecture for Text Engineering. *Journal of Computers and the Humanities*, 36(2): 223–254.
- Diana, M., Cunningham, H., Bontcheva, K., Catizone, R., Demetriou, G., Gaizauskas, R., Hamza, O., Hepple, M., Herring, P., Mitchell, B., Oakes, M., Peters, W., Setzer, A., Stevenson, M., Tablan, V., Ursu, C. & Wilks, Y. 2000. A Survey of Uses of GATE. Laporan Teknikal. Department of Computer Science, University of Sheffield.
- Diana, M., Tablan, V., Ursu, C., Cunningham, H. & Wilks, Y. 2001. Named Entity Recognition from Diverse Text Types. *Recent Advances in Natural Language Processing Conference*, hlm. 257–274.
- Erk, K. and Pado, S. 2005. Analyzing Models For Semantic Role Assignment Using Confusability. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, Canada*, hlm. 668–675.
- Fazilah. J. 2006. *Aplikasi Peranan Semantik Dalam Pemprosesan Bahas Tabii*. Tesis Sarjana Muda, Jabatan Sains Komputer, Universiti Kebangsaan Malaysia.
- Fillmore, C. J. 1968. "The Case for Case". IN *Universals in Linguistic Theory*. Bach. E. and Harms, R. (eds.). New York, hlm. 1-88.
- Friedl, J. E. F. 1997. *Mastering Regular Expression*. O'Reilly and Associates, USA.
- Gildea, D and Jurafsky. 2000. D. Automatic Labeling of Semantic Roles. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong*, hlm. 512 – 520.
- Hacioglu, K. 2004. A Lightweight Semantic Chunking Model Based on Tagging. *Proceeding of HLT*.
- Hepple, M. 2000. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, hlm. 278-285.
- Hockenmaier, J. & Steedman, M. 2002. Generative Models for Statistical Parsing With Combinatory Grammars. *Proceedings of the ACL*, hlm. 335–342.
- Jackendoff, R. 1990. *Semantic Structure*. London Massachusetts Institute of Technology, MIT Press.
- Jurafsky, D. & Martin, J.H. , 2000. *Speech and Language Processing*, Prentice Hall, New Jersey.
- Kamus Dewan. 2005. Ed. 4. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Lin, D. 1998. Dependency-based evaluation of MINIPAR. Workshop on the Evaluation of Parsing Systems, hlm. 234-241.
- Liu, R.L. & Soo, V. 1993. An Emperical Study on Thematic Knowledge Acquisition Based on Syntatic Clues and Heuristics. *Proceeding. of 31st Annual Meeting of the ACL*, Columbus, Ohio, hlm. 243-250.
- Tajularifin. 2007. Kaedah-Kaedah Heuristik. Lecture Notes in Asas Pendidikan. Universiti Putra Malaysia.pp.24-25.
- Omar, N, Hanna, P, Kevitt, P.M. 2004. Heuristics-based entity-relationship modeling through natural language processing. *Proceeding of The Fifteenth Irish Conference on Artificial Intelligence and Cognitive Science Department*.
- Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., & Jurafsky, D. 2005. Semantic Role Labeling Using Different Syntactic Views. In *Proceedings of the 43rd Annual Meeting of the ACL*, hlm. 581-588.
- Punyakanok, V., Roth, D., Yih, W., & Zimak, D. 2004. Semantic Role Labeling Via Integer Linear Programming Inference. *Proceedings of the 20th International Conference On Computational Linguistics*, hlm. 1346-1352.
- Rijsbergen, C. V. 1979. *Information Retrieval*. USA: Butterworth-Heinemann.
- Sundheim, B. 1995. Overview of Results of the MUC-6 Evaluation. *Proceedings of the Sixth Message Understanding Conference*, hlm. 13-32.
- Swier. R and S. Stevenson. 2004. Unsupervised Semantic Role Labeling. In *Proceedings EMNLP'04*, Barcelona,
- Schank and Roger C. 1972. Conceptual dependency.A theory of natural language understanding. *Journal Of Cognitive Psychology*, hlm. 3:552-63.
- Xpress- MP. 2003. Dash Optimization. Xpress- MP. <http://www.dashoptimization.com/products.html>. [14 Disember 2007]

Zanakis, S.H, Evan, J.R. 1981. *Heuristic "Optimization": Why,When and How to Use It*. Interfaces Vol. 11, No. 5.

Nazlia Omar
Jabatan Sains Komputer
Fakulti Teknologi dan Sains Maklumat
43000 UKM Bangi, Selangor
no@ftsm.ukm.my

Siti Salwa Hasbullah
MIMOS Berhad
Technology Park Malaysia
57000 Kuala Lumpur
Salwa_hasbullah@yahoo.com