PENGECAMAN ENTITI NAMA BAHASA MELAYU BERASASKAN PERATURAN

ULFA NADIA

UNIVERSITI KEBANGSAAN MALAYSIA

PENGECAMAN ENTITI NAMA BAHASA MELAYU BERASASKAN PERATURAN

ULFA NADIA

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGAKUAN

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

07 Jun 2018

ULFA NADIA
GP04491

PENGHARGAAN

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

Ucapan terima kasih yang tidak terhingga juga ditujukan kepada ahli keluarga tercinta khususnya ibu saya Zuraidah, bapa saya Ir. Chairul Nursi, adik saya Masfidia Ridifa, Muhammad Ziad Alfarazi serta Muhammad Hafizul Qura atas segala doa, pengorbanan, dorongan dan kesabaran yang diberikan sepanjang saya menyiapkan kajian ini. Ucapan penghargaan ini juga ditujukan kepada kakak seperjuangan semasa di FTSM, Fanindia Purnama Sari, M.IT yang banyak memberikan tunjuk ajar, nasihat dan motivasi dalam mengharungi cabaran-cabaran sepanjang pengajian ini.

Akhir bicara, saya mengucapkan terima kasih kepada mereka yang terlibat secara langsung atau tidak langsung sehingga terhasilnya tesis ini.

ABSTRAK

Pengecaman Entiti Nama (PEN) merupakan satu proses menganotasi atau memberi tanda nama dalam ayat untuk setiap kelas perkataan seperti nama individu, organisasi, tarikh, masa, dan lain-lain. Proses ini merupakan satu proses penting daripada sebahagian tugas asas dalam capaian maklumat, bagi prestasi pemprosesan teks. Masalah utama dalam PEN bahasa Melayu ialah penggunaan entiti nama yang mempunyai rujukan silang dengan entiti nama lain, pencampuran entiti nama yang berbeza dan pengulangan entiti nama. Objektif utama kajian ini adalah untuk membangun peraturan baru bagi PEN bahasa Melayu dan membandingkan prestasi PEN bahasa Melayu berasaskan peraturan dengan kajian lepas. Proses ini bermula dengan penyediaan korpus, pembangunan gazetir, pembangunan peraturan dan penilaian. Penyediaan korpus menggunakan fail teks yang diperoleh daripada berita atas talian meliputi pelbagai domain. Sebanyak 200 korpus telah dipilih dan 170 daripadanya dijadikan sebagai korpus latihan manakala baki selebihnya sebagai korpus ujian. Seterusnya data latihan diguna pada proses pembangunan gazetir dan proses pembangunan peraturan. Pembangunan peraturan merupakan proses yang melibatkan senarai gazetir. Pembangunan peraturan entiti nama dalam kajian ini memberi fokus kepada pengecaman entiti nama yang melibatkan 9 entiti iaitu nama individu, lokasi, organisasi, jawatan, tarikh, masa, kewangan, ukuran dan peratusan. Proses penilaian pula dilaksanakan bagi melihat keberkesanan peraturan PEN yang dibangunkan serta membuat perbandingan hasil PEN dengan kajian lepas. Secara keseluruhannya, pengujian ini memberikan hasil dengan nilai kejituan 90.23%, dapatan 92.13% dan ukuran-f 91.05% berbanding prestasi kajian lepas iaitu 94.44% nilai dapatan, 85% kejituan dan 89.47% ukuran-f. Hasil daripada kajian ini diharap dapat membantu penyelidik dalam melaksanakan PEN bahasa Melayu dengan menghasilkan nilai ketepatan yang lebih tinggi.

MALAY NAMED ENTITY RECONGNITION USING RULED-BASED APPROACH

ABSTRACT

Name Entity Recognition (NER) is a process of recognition or label in a sentence for each word phrase as person name, organization name, date, time, and so on. This process is an important as basis of information retrieval for text processing performance. The main problem in Malay NER is the use of named entities that have cross-references with other entities, combining different names of entities and repeating named entities. The main objective of this study is to develop new rules for Malay NER and compare the performance of rule based approach for NER Malay with the previous study. This process begins with selecting the corpus, developing gazetteer, developing rules and evaluation. The provision of the corpus uses text files, obtained from online over various domains. The total number of corpus selected is 200, in which 170 of them have been used as a training corpus whiles the rest as a test corpus. Then, the training corpus is selected for generating the gazetteer and developing rules. The rule development is a NER process by involving gazetteer. This study focuses on nine entities, i.e person, location, organization, position, date, time, currency, measurement and percentage. The evaluation process uses precision, recall and f-measure and compares the results of the NER with the previous study. Overall, the evaluation shows a precision of 90.23%, recall 92.13% and f-measure 91.05% compared to the previous study of 85% precision, 94.44% recall and 89.47% f-measure. Outcome from this research is expected to help other researchers in implementing the Malay NER using rule based approach through the addition of new rules to achieve higher accuracy.

KANDUNGAN

Halaman

PENGAKUAN		**ii**
PENGHARGAAN		**iii**
ABSTRAK		**iv**
ABSTRACT		**v**
KANDUNGAN		**vi**
SENARAI JADUAL		**viii**
SENARAI ILUSTRASI		**ix**
SENARAI SINGKATAN		**x**
BAB I	**PENDAHULUAN**	**1**
1.1	Pengenalan	1
1.2	Latar Belakang Kajian	2
1.3	Penyataan Masalah	4
1.4	Matlamat Dan Objektif Kajian	5
1.5	Skop Kajian	6
1.6	Metod Kajian	6
1.7	Organisasi Kajian	7
1.8	Kesimpulan	8
BAB II	**KAJIAN SUSASTERA**	**9**
2.1	Pengenalan	9
2.2	Pengecaman Entiti Nama	9
2.3	Kaedah Pengecaman Entiti Nama	11
	2.3.1 Pendekatan Berasaskan Peraturan	12
	2.3.2 Pendekatan Sistem Berselia	13
	2.3.3 Pendekatan Sistem Separa Berselia	14
	2.3.4 Pendekatan Sistem Tanpa Selia	15
	2.3.5 Perbandingan Pendekatan Berdasarkan Peraturan dan Statistik	15
2.4	Kajian Lepas	16
	2.4.1 PEN Selain Bahasa Melayu	16
	2.4.2 PEN Bahasa Melayu	18
2.5	Kesimpulan	25

BAB III	**METODOLOGI**	**26**
3.1	Pengenalan	26
3.2	Fasa Metodologi Penyelidikan	26
3.3	Fasa Penyediaan Korpus	29
3.4	Fasa Pembangunan Gazetir	30
3.5	Fasa Pembangunan Peraturan	31
	3.5.1 Analisis Beberapa Peraturan	32
	3.5.2 Pembangunan Peraturan	34
3.6	Fasa Penilaian	47
3.7	Kesimpulan	47
BAB IV	**IMPLIMENTASI DAN PERBINCANGAN**	**48**
4.1	Pengenalan	48
4.2	Penilaian Korpus	48
	4.2.1 Korpus Latihan	48
	4.2.2 Korpus Ujian	50
4.3	Analisis Keputusan	51
4.4	Perbandingan Penilaian	54
4.5	Kesimpulan	56
BAB V	**KESIMPULAN**	**57**
5.1	Pengenalan	57
5.2	Rumusan Kajian	57
5.3	Sumbangan Kajian	59
5.4	Kekangan Kajian	60
5.5	Cadangan Penambahbaikan	60
5.6	Penutup	61
RUJUKAN		**62**
LAMPIRAN		
Lampiran A	Senarai Gazetir	67

SENARAI JADUAL

No. Jadual		Halaman
Jadual 2.1	Perbandingan pendekatan berasaskan peraturan dan statistik	16
Jadual 2.2	Perbandingan kajian lepas PEN bahasa Melayu	22
Jadual 3.1	Aktiviti setiap fasa	29
Jadual 3.2	Entiti nama	31
Jadual 3.3	Taburan entiti nama	35
Jadual 3.4	Ungkapan nalar dalam pembangunan peraturan	36
Jadual 3.5	Peraturan PEN bahasa Melayu	37
Jadual 4.1	Contoh artikel data latihan	49
Jadual 4.2	Hasil keputusan bagi masing-masing entiti	51
Jadual 4.3	Hasil keputusan korpus Alfred et al. (2014)	55
Jadual 4.4	Perbandingan keputusan antara dua PEN	56

SENARAI ILUSTRASI

No. Rajah		Halaman
Rajah 3.1	Aliran metodologi penyelidikan	27
Rajah 3.2	Simulasi PEN	28
Rajah 3.3	Contoh korpus latihan hasil anotasi manual	35
Rajah 3.4	Pembangunan peraturan entiti nama	36
Rajah 4.1	Sebahagian entiti nama data latihan	50
Rajah 4.2	Peratusan ketepatan PEN	52

SENARAI SINGKATAN

KNK	Kata Nama Khas
MEMM	Model Entropi Maksimum Markov
MMT	Model Markov Tersembunyi
PBT	Pemprosesan Bahasa Tabii
PEN	Pengecaman Entiti Nama
PM	Pembelajaran Mesin

BAB I

PENDAHULUAN

1.1 PENGENALAN

Pemprosesan Bahasa Tabii (PBT) atau *Natural Language Processing* (NLP) merupakan salah satu cabang ilmu AI (*Artificial Inteligence*) yang berfokus pada pengolahan bahasa tabii yang berkaitan dengan sistem komputer yang memahami dan menganalisa bahasa yang digunakan oleh manusia (Budi et al. 2005). PBT ialah salah satu perintisnya aspek komunikasi manusia-mesin kerana bahasa merupakan salah satu keperluan asas komunikasi manusia. Objektif utama PBT adalah untuk membangunkan model untuk memproses tugas linguistik seperti membaca, menulis, mendengar dan bercakap (James 2003).

PBT dijalankan secara berperingkat pada penghujung tahun 1940an yang memberi tumpuan kepada penterjemahan mesin (*machine translation*). Penyelidikan dalam bidang ini seterusnya berkembang pada tahun 1990-an sehingga kini (Jiang et al. 2016). Pada masa kini, PBT melibatkan pemprosesan teks berstruktur dan tidak berstruktur. Teks berstruktur lebih mudah untuk dianalisis kerana disusun secara tetap atau coraknya telah diketahui, seperti teks yang ditandakan dalam format HTML atau XML. Sebaliknya, menganalisis teks tidak berstruktur sulit kerana ciri bentuk yang dinamik dan bebas (Jones 2001).

Terdapat beberapa cabang tugas utama atau kajian yang lazimnya dilaksanakan dalam bidang PBT selain penterjemahan mesin, seperti capaian maklumat (*information retrieval*), pengecaman entiti nama (*named entity recognition*) dan pengekstrakan maklumat (*information extraction*) (Oudah et al. 2016). Pengecaman Entiti Nama (PEN) didefinisikan sebagai proses mengenalpasti entiti nama dan mengklasifikasikannya ke dalam kategori tertentu (Thenmalar et al. 2015).

Sebagai sebahagian daripada tugas asas dalam PBT, Pengecaman Entiti Nama (PEN) memerlukan pengenalpastian nama-nama yang sepatutnya dari data yang tidak tersusun dan mengkelaskannya dalam kategori yang telah ditentukan. Kategori seperti nama orang, nama organisasi, tarikh, masa, dan lain-lain, digolongkan dibawah tiga kelas diperingkat pertama: Entiti Nama (ENAMEX), Numerik (NUMEX) dan Masa (TIMEX). Ciri-ciri kelas yang berbeza ini menjadikannya lebih mudah untuk diklasifikasikan ke dalam subkelas lanjut (Aboaoga & Aziz 2013). Pendekatan serupa juga digunakan dalam sistem yang dicadangkan ini.

Sebelum ini, terdapat banyak kajian penyelidikan telah dilakukan dalam bidang PEN yang melibatkan pelbagai bahasa selain bahasa Melayu, antaranya bahasa Inggeris-China (Druzhkina & Stepanova 2016), Jerman (Druzhkina & Stepanova 2016), Bahasa Turki (Druzhkina & Stepanova 2016), Bahasa Arab (Karaa & Slimani 2017), Bahasa Kanada (Pallavi et al. 2018), Urdu (Riaz 2010) dan Bahasa Indonesia (Aryoyudanta et al. 2017; Leonandya et al. 2016; Wibawa & Purwarianti 2016). Teknik yang digunakan ialah Entropi Maksimum (EM), Model Markov Tersembunyi (MMT), *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM).

1.2 LATAR BELAKANG KAJIAN

Peningkatan dokumen elektronik dari tahun ke tahun (Kanimozhi & Venkatesan 2015) menjadikan penyelidikan dalam Pemprosesan Bahasa Tabi'i (PBT) semakin penting dalam capaian semula maklumat (Jiang et al. 2016). PBT mempunyai potensi yang tinggi dalam kebanyakan aplikasi seperti rumusan artikel sistem soal jawab, sistem penilaian sebut harga dan sistem tutorial (Pallavi et al. 2018). PBT menjadikan maklumat tidak berstruktur kepada satu maklumat yang lebih bermakna. PBT boleh memproses artikel ke dalam bentuk yang lebih jelas dengan menghubungkan satu perkataan kepada perkataan yang lain. Antara cabang penyelidikan PBT ialah Pengecaman Entiti Nama (PEN).

PEN merujuk pada proses mencari sebahagian daripada teks yang mewakili kata nama khas dan kemudian mengklasifikasikannya dalam kategori yang sesuai. Sebahagian daripada teks tersebut boleh jadi sebagai indeks, link, dan dapat diguna juga dalam sistem soal jawab (Sazali 2016). PEN melakukan tugas tidak hanya

mengenalpasti entiti nama dalam teks yang tidak terstruktur namun juga mengklasifikasikannya mengikut susunan jenis entiti yang telah ditentukan. Tugas PEN pertama kali dilakukan semasa persidangan MUC-6 ialah untuk menemukan jenis entiti, seperti orang, lokasi, dan organisasi serta masa, mata wang, dan peratusan dalam teks tidak terstruktur (Yong et al. 2011). Sebagai contoh terdapat nama individu dalam ayat *"Tuanku Syed Faizuddin Putra Jamalullail berceramah di Merlimau, petang ini…"*. Untuk mengklasifikasikan nama orang iaitu *"Tuanku Syed Faizuddin Putra Jamalullail"* ke dalam entiti nama individu dan "*Merlimau*" ke dalam entiti lokasi maka terdapat tiga pendekatan teknik yang perlu dilakukan seperti pendekatan berasaskan peraturan, pendekatan berasaskan statistik serta gabungan kedua-dua pendekatan ini (Alfred et al. 2014).

Pendekatan berasaskan peraturan ialah peraturan yang dibangun oleh ahli linguistik untuk mengenalpasti entiti nama daripada segi morfologi, sintatik atau memerlukan kata kunci yang mencerminkan sifat-sifat teks (Aboaoga & Aziz 2013). Pendekatan ini menggunakan corak yang dibuat secara manual kepada perkataan dalam ayat dengan menggunakan satu set peraturan yang ditulis. Selain daripada mudah digunakan kerana ia menggunakan peraturan yang ringkas dan menjimatkan ruang (Mubarak et al. 2015). Penambahbaikan ke atas PEN juga mudah dilakukan kerana pendekatan ini menggunakan set kecil peraturan yang mudah dan kurang kompleks (Al-Olimat et al. 2017). Diantara contoh pendekatan berasaskan peraturan dalam PEN ialah "jika sesebuah kata nama khas yang memiliki ciri-ciri nama orang seperti adanya gelaran maka itu ialah termasuk entiti nama individu". Manakala statistik pula melibatkan penggunaan teknik pembelajaran mesin yang dibagi dalam 3 jenis iaitu dengan berselia, separa berselia dan tanpa selia (Alfred et al. 2014; Art et al. 2015; Sharum & Abdullah 2011).

Penyelidikan mengenai PEN telah berkembang, satu di antara bahasa dunia yang banyak digunakan seperti berbahasa Inggeris kerana korpus yang sudah tersedia sehingga memudahkan penyelidikan dibidang PEN. Manakala korpus bahasa Melayu tidak sebanyak sumber bahasa Inggeris dan tiada koleksi yang boleh diguna bagi PEN (Sazali 2016). Hal ini berlaku kerana sumber bahasa Inggeris tidak dapat diguna dalam bahasa Melayu. Jika PEN bahasa Melayu menggunakan korpus bahasa Inggeris akan

memerlukan proses terjemah bahasa sehingga perlu masa yang lama untuk menyelesaikan pengecaman entiti nama. Sulaiman et al. (2017) mengecam entiti nama bahasa Melayu mengguna dua sistem yang sedia ada iaitu Stanford dan Illnois yang mengguna bahasa Inggeris. Hasil kajian menunjukkan terdapat kesalahan kerana perbezaan morfologi antara bahasa Inggeris dengan bahasa Melayu sehingga memperolehi ketepatan yang rendah. Namun PEN pada bahasa Melayu memiliki kesamaan dengan bahasa Inggeris seperti penggunaan huruf besar dan kata nama khas untuk mengecam entiti (Anthony 2013).

Satu fenomena yang berlaku dalam PEN bahasa Melayu ialah domain tatabahasa seperti morfologi dan konteks (Awang 2010). Terdapat dua bahagian dalam tatabahasa iaitu konteks yang merupakan bahagian pembentukan frasa dan ayat, dan morfologi iaitu satu bidang ilmu yang mengkaji perkataan dari segi struktur, bentuk dan penggolongan kata. Sebagai bahasa yang kaya dengan kosa katanya, bahasa Melayu mempunyai morfologinya yang tersendiri bagi menghasilkan perkataan lain (Mohamed et al. 2011). Berdasarkan kajian lepas PEN bahasa Melayu pada pendekatan berasaskan statistik didapati hasil keputusan yang rendah jika dibandingkan dengan pendekatan berasaskan peraturan dan gazetir (Anthony 2013). Oleh itu, PEN bahasa Melayu dalam kajian ini dengan menggunakan pendekatan berasaskan peraturan dan gazetir perlu dibangunkan.

1.3 PENYATAAN MASALAH

Beberapa kajian PEN berdasarkan peraturan telah dilakukan untuk pelbagai bahasa, tetapi masih sedikit kajian PEN dalam bahasa Melayu kerana korpus yang terhad (Morsidi et al. 2017). Antara masalah PEN ialah kehadiran entiti nama bersama perkataan *dan* dan simbol. Contoh *"Kementerian Perdagangan Dalam Negeri, Koperasi dan Kepenggunaan"*. Pada ayat tersebut terdapat simbol koma (,) yang dapat memisahkan perkataan lainnya atau merupakan entiti yang sama dari ayat tersebut. Alfred et al. (2014) menggunakan kamus yang mengandungi senarai entiti yang mengandungi perkataan *dan*, simbol koma dan simbol & akan tetapi kamus yang dibangunkan tidak mampu menghadapi perubahan pada entiti nama.

Selain itu, penggunaan perkataan *dan,* simbol koma dan simbol & boleh mengakibatkan ralat dalam mengenalpasti dua entiti yang berbeza. Contohnya pada ayat *"Kementerian Pendidikan"* dan *"Kementerian Kesihatan"* merupakan dua entiti yang berbeza. Antara masalah lainnya ialah penggunaan entiti nama lokasi yang mempunyai rujukan silang dengan entiti nama lain seperti individu dan lokasi. Contohnya perkataan *"Hospital Tun Aminah"* yang dapat memisahkan entiti *"Tun Aminah"* dalam entiti organisasi menjadi entiti individu. Manakala perkataan "Medan anak Nuying" juga memisahkan kata "Medan" dalam entiti lokasi (Alfred et al. 2014; Kamil 2015).

Masalah lain ialah pencampuran entiti nama yang berbeza. Sebagai contoh *"Pengarah Urusan U-Mobile Encik Muhammad bin Kassim"*. Dalam ayat tersebut terdapat tiga entiti yang berbeza iaitu *"Pengarah Urusan"*, '*U-Mobile*' dan *"Encik Muhammad bin Kassim*". Bagi manusia, tugas meletakkan tag bagi ketiga-tiga ayat tersebut adalah mudah kerana manusia mempunyai akal pengetahuan dan naluri yang boleh memproses maklumat dan dapat mengenalpasti perbezaan makna perkataan yang digunakan pada ayat tersebut (Kumawat & Jain 2015). Namun, ia agak sukar bagi sistem PEN terutamanya apabila tiada peraturan khusus untuk perkara ini.

Sementara itu, pengulangan entiti dengan nama pendek juga berlaku dalam banyak artikel. Sebagai contoh *"Tan Sri Zulhasnan Rafique, 62, sudah lapan bulan... Di Amerika Syarikat, komen negatif merupakan perkara biasa," kata Zulhasnan"*. Pada contoh tersebut perkataan *"Zulhasnan"* merupakan sebagian nama daripada "*Tan Sri Zulhasnan Rafique*". Hal ini juga masih menjadi masalah dalam PEN (Alfred et al. 2014; Kamil 2015).

Kekurangan peraturan dalam PEN bahasa Melayu dalam kajian sedia ada menyebabkan timbulnya berbagai masalah yang belum dapat diselesaikan. Oleh yang demikian pembangunan peraturan baru dapat mengurangkan ralat dalam PEN.

1.4 MATLAMAT DAN OBJEKTIF KAJIAN

Objektif kajian ini adalah seperti berikut:

i. Membina peraturan baru bagi PEN bahasa Melayu.

ii. Membandingkan prestasi PEN bahasa Melayu berasaskan peraturan dengan kajian lepas.

1.5 SKOP KAJIAN

Skop kajian ini tertumpu kepada sumber berita dalam bahasa Melayu yang dipetik daripada berita harian atas talian yang mencakupi semua bidang, terutamanya seperti Bernama, Berita Harian (BH) dan Malaysia Kini.

1.6 METOD KAJIAN

Metod kajian yang digunakan dalam kajian ini merangkumi metodologi penyelidikan yang memaparkan ringkasan keseluruhan proses kajian yang dijalankan bermula daripada proses penyediaan korpus, pembangunan gazetir, pembangunan peraturan dan penilaian.

Proses metodologi penyelidikan menerangkan dengan lebih terperinci tentang proses pembangunan kerangka kerja PEN berdasarkan peraturan. Kerangka kerja konsep ini merupakan garis panduan dalam menyelesaikan PEN dalam bahasa Melayu

Proses penyediaan korpus yang dijalankan dari awal permulaan kajian dengan menggunakan data sekunder yang dipetik daripada berita harian yang didapati secara atas talian. Sebanyak 200 korpus telah dipilih dan 170 daripadanya dijadikan sebagai korpus latihan semasa pembangunan peraturan manakala selebihnya dijadikan sebagai korpus ujian semasa proses penilaian.

Proses pembangunan gazetir menerangkan pembangunan senarai sebahagian entiti nama secara manual bagi membangun peraturan PEN. Dalam proses ini gazetir bagi beberapa PEN telah dibina untuk menentukan entiti nama yang sesuai kategori.

Proses pembangunan peraturan pula merupakan proses bagi membangun peraturan PEN. Sebanyak 9 entiti nama dan masing-masing terdapat 1-5 peraturan dibangunkan dalam proses ini. Peraturan bagi setiap PEN turut dibangunkan mengikut

aturan masing-masing. Aturan ini amat penting kerana ia memberi kesan terhadap hasil penandaan PEN pada sesi pengujian kelak.

Proses terakhir merupakan proses penilaian bagi melihat keberkesanan peraturan yang telah dibangunkan bagi menghasilkan keputusan PEN. Hasil pengujian pembangunan peraturan yang diukur melalui kejituan PEN ini dibandingkan dengan hasil yang diperolehi daripada PEN kajian lepas.

1.7 ORGANISASI KAJIAN

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

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

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

Bab III pula memperincikan metod kajian yang digunakan dalam kajian ini yang terdiri daripada metodologi penyelidikan, penyediaan korpus, pembangunan gazetir, pembangunan peraturan dan penilaian. Pembangunan peraturan bagi 9 jenis PEN juga diperincikan dalam bab ini.

Bab IV merupakan fasa implimentasi dan perbincangan di mana pengujian dinilai dari segi kejituan, dapatan dan ukuran-f seterusnya membandingkan ketepatan antara kajian ini dan kajian sebelumnya dalam Pengecaman Entiti Nama berdasarkan peraturan yang dibangunkan. Analisis keputusan secara terperinci turut dilakukan dalam bab ini.

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

1.8 KESIMPULAN

Bab ini menerangkan secara menyeluruh tentang latar belakang kajian, pernyataan masalah tentang skenario semasa yang sedang dihadapi, matlamat dan objektif kajian, skop kajian serta penerangan ringkas tentang metod kajian yang digunakan. Objektif bagi kajian ini adalah untuk menambahbaik peraturan bagi pengecaman entiti nama bahasa Melayu berdasarkan kajian lepas sedia ada.

BAB II

KAJIAN SUSASTERA

2.1 PENGENALAN

Bab ini membincang hasil tinjauan kajian susastera terhadap beberapa kajian yang berkaitan dengan pengecaman entiti nama yang telah dijalankan oleh penyelidik terdahulu. Secara amnya, kajian ini dimulai dengan pengenalan Pengecaman Entiti Nama (PEN) dan jenisnya. Kemudian berkaitan dengan kajian lepas yang dijalankan dalam bidang PEN. Kaedah yang dibincang akan memberikan tumpuan khusus kepada analisis kandungan dokumen yang dihasilkan oleh teks berita. Justifikasi pemilihan kajian diterangkan pada bahagian seterusnya dan kesimpulan pada akhir bab ini.

2.2 PENGECAMAN ENTITI NAMA

Pengecaman entiti nama atau PEN merupakan elemen dasar dan memainkan peranan penting khususnya dalam bidang yang berkaitan dengan Pemprosesan Bahasa Tabii (PBT). Pengecaman entiti nama memiliki tugasan untuk mengenalpasti, mengecam dan mengklasifikasi terma atau frasa mengikut entiti nama daripada teks yang tidak terstruktur. PEN menjadi sub tugas ekstraksi informasi yang melibatkan kata nama khas dalam teks yang disebut entiti nama dan kelas dari entiti nama yang telah ditentukan sebelumnya, misalnya nama individu, lokasi dan organisasi. Maklamat utama daripada PEN ialah untuk mengurangi penandaan manual entiti nama dalam teks yang banyak mengambil masa dan proses yang panjang (Ulanganathan et al. 2017).

Perbezaan ciri merupakan asas untuk mengecam entiti nama. Hal ini dapat diperhatikan dengan kriteria ortografi yang ditandai oleh penggunaan huruf besar di awal perkataan. PEN pertama kali diperkenalkan pada mesyuarat *The Sixth Message Understanding Conference (MUC-6)* yang menghasilkan suatu keputusan untuk membahagi entiti nama menjadi ENAMEX merujuk kepada orang, organisasi dan lokasi, TIMEX merujuk kepada tarikh dan masa, dan NUMEX merujuk kepada mata wang, peratus dan kuantiti (Budi et al. 2005). Selain itu, pelabelan entiti nama untuk sesebuah objek juga boleh diubah suai mengikut keperluan domain tertentu. Dataset yang diperkenal pada mesyuarat MUC7 adalah sebahagian daripada *North American News Text Corpora* yang dianotasi dengan pelbagai jenis entiti termasuk orang, lokasi, organisasi, peristiwa temporal, unit kewangan, dan sebagainya. Oleh kerana tiada pemetaan langsung dari peristiwa temporal, unit kewangan, dan entiti lain dari MUC7 dan label MISC dalam dataset CoNLL03, biasanya pengecaman entiti nama hanya mengenali entiti seperti orang, organisasi dan lokasi (Yong et al. 2011).

Awalnya, teknik PEN banyak diaplikasikan dalam bahasa Inggeris. Kemudian pada tahun 2002, satu persidangan *Conference on Computational Natural Language Learning (CoNNL-2002)* dianjurkan bertujuan untuk memberi fokus pengecaman entiti nama dalam bahasa selain bahasa Inggeris. Beberapa penyelidikan telah dijalankan menggunakan bahasa selain bahasa Inggeris seperti bahasa Indonesia (Budi et al. 2005; Wibawa & Purwarianti 2016), bahasa Arab (Aboaoga & Aziz 2013) dan bahasa Melayu (Bali et al. 2007; Sharum et al. 2011; Alfred et al. 2013; Sulaiman 2017; Ulanganathan et al. 2017). Setiap bahasa memiliki keunikan sendiri berbanding dengan bahasa lain, sehingga memerlukan peraturan yang berbeza untuk menentukan suatu entiti nama dari perkataan.

Ada beberapa elemen yang perlu diketahui dalam pengecaman entiti nama, dari perkataan tunggal hingga perkataan tersebut menjadi suatu perkataan yang kompleks membentuk sebuah ayat. Ciri dapat dikenalpasti oleh beberapa kategori iaitu atribut Boolean bernilai benar jika mengandungi huruf besar dan sebaliknya. Atribut numerik merujuk kepada panjang karakter sebuah perkataan dan atribut nominal merujuk kepada huruf kecil dari sesuatu perkataan. Terdapat dua jenis ciri yang biasa dikenali untuk mengecam entiti nama iaitu ciri aras perkataan dan *list lookup features* (Morsidi et al. 2017).

Ciri aras perkataan merupakan ciri perkataan yang tidak bergantung kepada konteks tertentu (Wibawa & Purwarianti 2016) sedangkan *list lookup features* disebut sebagai perkataan yang telah ditentukan dalam senarai yang dikenali sebagai gazetir. Istilah gazetir, leksikon, dan kamus memiliki takrifan yang sama dengan senarai. Ciri senarai menandakan terdapat hubungan antara entiti dalam senarai, contoh Kuala Lumpur ialah bandar. Jika perkataan *Kuala* terdapat pada senarai, maka kebarangkalian perkataan ini terdapat pada ayat adalah tinggi.

Dalam bidang kajian linguistik, terdapat tujuh kriteria yang digunakan bagi mengkategorikan perkataan ke dalam sesuatu golongan, iaitu kriteria fonologi, morfologi, sintaksis, leksikal, semantik, pragmatik dan wacana (*discourse*) (Yunus et al. 2010). Namun begitu, kajian ini hanya memfokuskan kepada morfologi dan sintaksis sahaja memandangkan kedua-dua kriteria tersebut bertepatan dengan kehendak kajian yang dijalankan ini.

Morfologi merupakan kajian mengenai struktur, bentuk dan penggolongan kata. Salah satu struktur morfologi ialah penggunaan huruf besar (Nadeau & Sekine 2006; Ranaivo-Malangon et al. 2015). Sintaksis pula merupakan istilah bagi cara susunan dan urutan dalam ayat. Ianya memerlukan tatabahasa dan penghurai (bagi menghuraikan perkataan dan frasa kepada beberapa bahagian untuk memahami maksud perkataan dan hubungan antara perkataan dalam ayat) yang memberi tumpuan kepada analisis perkataan dalam ayat bagi mempamerkan struktur tatabahasa dalam ayat tersebut (Yunus et al. 2010). Dalam struktur morfologi bahasa Melayu, penggunaan huruf besar pada awal perkataan menunjukkan sebuah perkataan adalah kata nama khas. Hal ini memudahkan untuk mengenal pasti sama ada perkataan tersebut menunjukkan entiti nama atau bukan (Morsidi et al. 2017).

2.3 KAEDAH PENGECAMAN ENTITI NAMA

Untuk mengautomasikan proses pengecaman entiti nama, enjin perlu dilatih agar dapat menganalisis dan memahami kandungan teks sebelum mengenalpasti entiti yang disebutkan (Ulanganathan et al. 2017). Beberapa kajian lepas menunjukkan bahawa Pengecaman Entiti Nama (PEN) boleh dilaksanakan melalui empat pendekatan iaitu pendekatan berasaskan peraturan (*rule-based approach*), pendekatan sistem yang

berselia (*supervised*), pendekatan separa selia (*semi supervised*) dan pendekatan sistem tanpa berselia (*unsupervised*). Setiap pendekatan ini mempunyai kelebihan dan kekurangan yang tersendiri berdasarkan saiz korpus dan domain yang dipilih (Morsidi et al. 2015).

2.3.1 Pendekatan Berasaskan Peraturan

Pengenalan entiti nama berasaskan peraturan ialah pendekatan menggunakan satu set peraturan yang disedia (Kumawat & Jain 2015). Peraturan ini dibangun bagi upaya mengatasi kurangnya sumber daya korpus untuk bahasa tertentu (Aboaoga & Aziz 2013). Kaedah ini mengesan entiti nama dengan menyusun satu set aturan yang dibuat secara manual merujuk kepada kamus yang telah dibangun sebelumnya (Alfred et al. 2014; Anthony 2013; Anthony et al. 2014; Art et al. 2015).

Sebahagian model pengecaman entiti nama menggunakan kamus sebagai sumber rujukan untuk menentukan entiti nama berdasarkan peraturan untuk menghasilkan prestasi yang tinggi dan optimal, sebahagian lagi menggunakan sebahagian kecil gazetir untuk menghasilkan nilai dapatan dan nilai kejituan yang tinggi. Pendekatan berdasarkan peraturan dalam mengenal pasti entiti nama bermaksud mengguna satu set peraturan yang terdahulu dan senarai kamus perkataan yang dibuat oleh manusia secara manual (Alfred et al. 2014).

Pada pendekatan berdasarkan peraturan, umumnya pembentukan perkataan dicapai berdasarkan morfologi dan struktur sintaks bahasa tertentu seperti menggunakan penandaan golongan kata, *word precedence*, ciri orthografik seperti penggunaan huruf besar dan kombinasi dengan menggunakan kamus (Budi et al. 2005). Pembangun sistem perlu mempunyai pengalaman dan kemahiran pengetahuan bahasa dan tata bahasa dalam menentukan ketepatan pengecaman entiti nama. Hal ini kerana proses pembangunan peraturan memerlukan pengujian yang dilakukan berulang kali bagi mendapatkan hasil pengecaman yang lebih tepat (Nadeau & Sekine 2006; Ranaivo-Malangon et al. 2015).

Budi et al. (2005) telah membangun satu model pengecaman entiti nama bagi bahasa Indonesia yang diberi nama InNER. Model ini menggunakan pendekatan

berdasarkan peraturan terhadap kontekstual, morfologi dan penandaan golongan kata (POS *tagger*) bagi mengenalpasti entiti dalam bahasa Indonesia. Model ini digunakan untuk mencari entiti nama orang dalam sebuah ayat.

Morsidi et al. (2017) menggunakan *regex* untuk mengesan ciri istimewa daripada kata sendi nama dengan mengenal pasti struktur kata benda pada perkataan dalam suatu perenggan. Penyelidikan ini terhad berdasarkan anggapan bahawa huruf awalan perkataan kata nama khas dimula dengan huruf besar. Kata sendi nama diguna untuk mewakili kata benda sebagai objek, yang diguna bersama dengan morfologi atau anotasi perkataan seperti kata sifat dan preposisi (Abu Bakar et al. 2013).

2.3.2 Pendekatan Sistem Berselia

Pendekatan sistem berselia ini termasuk dalam pembelajaran mesin yang banyak digunakan dalam PEN ialah seperti *Hidden Markov Model (HMM), Conditional Random Field (CRF), Naïve Bayes* (NB), *Neural Networks,* SVM (Kanimozhi & Venkatesan 2015). Kebanyakan pendekatan ini memerlukan korpus beranotasi dalam kuantiti yang banyak sebagai data latihan. Pendekatan PEN dalam pembelajaran mesin yang akhir-akhir ini dilakukan ialah CRF adalah kaedah probabilistik yang popular berstruktur ramalan. Teknik ini telah diterapkan di beberapa domain termasuk bioinformatik, visi komputer dan pemprosesan teks. Che et al. (2013) menciptakan satu aplikasi skala besar CRF pertama untuk segmentasi frasa kata nama khas dalam teks. CRF rantaian linear telah digunakan untuk pelbagai masalah dalam PBT termasuk PEN. Dalam PEN semua label entiti nama bersifat bebas tetapi label entiti nama perkataan selepasnya bersifat terikat, misalnya Los Angeles menunjukkan lokasi, manakala Los Angeles Times menunjukkan organisasi (Ulanganathan et al. 2017). Salah satu cara membuat tidak terikat ini ialah mengatur output pembolehubah dalam rantai linear yang CRF (Salleh et al. 2017).

Pendekatan berselia menggunakan korpus latihan yang mengira kebarangkalian urutan yang boleh digunakan sebagai alternatif kepada pendekatan kekerapan perkataan (Kumawat & Jain 2015). Model pendekatan sistem berselia memerlukan korpus PEN untuk dapat mengesan jenis entiti yang sebenar pada perkataan atau frasa baru