https://doi.org/10.17576/apjitm-2019-0801-04

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 8 No. 1, June 2019: 37 – 47
e-ISSN: 2289-2192

MALAY NAMED ENTITY RECOGNITION USING RULE BASED APPROACH

ULFA NADIA
NAZLIA OMAR

ABSTRACT

Named Entity Recognition (NER) research based on rules is widely investigated and is used in various languages mainly English. However, the English NER rules are different from the Malay language due to different morphology. One of the challenging issues in Malay are cross reference between named entities and entity repetition. This paper proposes to solve these issues in Malay NER. This study starts by providing Malay online news corpus, gazeteer development, rules development and evaluation. This study focus on nine named entities i.e person, organization, position, date, time, currency, measurement and percentage. Overall the experimental result shows 90.23% precision, 92.13% recall and 91.05% f-measure. The outcome from this research is expected to help other researchers in implementing the Malay NER using rule based to achieve higher accuracy.

Keywords: named entity, Malay, Named Entity Recognition, rule based.

PENGECAMAN ENTITI NAMA BAHASA MELAYU MENGGUNAKAN PENDEKATAN BERASASKAN PERATURAN

ABSTRAK

Kajian mengenai pengecaman entiti nama (PEN) berdasarkan peraturan telah dijalankan secara meluas dan diguna melalui pelbagai bahasa terutama bahasa Inggeris. Namun, peraturan yang dihasilkan oleh bahasa Inggeris memiliki perbezaan dengan bahasa Melayu kerana perbezaan morfologi. Isu yang mencabar dalam PEN bahasa Melayu adalah rujukan silang antara satu entiti nama dengan entiti nama lainnya, pencampuran entiti nama dan pengulangan entiti nama. Kertas ini mencadang peraturan baru bagi mengatasi isu dalam PEN bahasa Melayu. Kajian bermula dengan menyedia korpus fail teks berita dalam talian bahasa Melayu, pembangunan gazetir, pembangunan peraturan dan penilaian. Kajian ini memberi fokus kepada pengecaman entiti nama yang melibatkan sembilan entiti nama iaitu nama individu, lokasi, organisasi, jawatan, tarikh, masa, kewangan, ukuran dan peratusan. Secara keseluruhannya, pengujian ini memberikan hasil dengan nilai kejituan 90.23%, dapatan 92.13% dan ukuran-f 91.05%. Hasil daripada kajian ini diharap dapat membantu penyelidik dalam melaksanakan PEN bahasa Melayu dengan menghasilkan nilai ketepatan yang lebih tinggi.

Kata Kunci: entiti nama, pengecaman entiti nama bahasa Melayu, peraturan.

PENGENALAN

Pemprosesan Bahasa Tabii (PBT) atau Natural Language Processing (NLP) adalah salah satu cabang ilmu AI (Artificial Inteligence) yang bertumpu pada pengolahan bahasa tabii. Lazimnya bahasa tabii adalah bahasa yang perlu difahami oleh komputer sebagai medium komunikasi antara mesin dan manusia yang hanya difahami oleh manusia (Halid & Omar 2017). Penyelidikan dalam bidang PBT berkembang pada tahun 1990-an sehingga kini (Suhaimi & Nazlia 2014). PBT dijalankan secara berperingkat pada penghujung tahun 1940an yang

memberi tumpuan kepada penterjemahan mesin (*machine translation*). Pada masa kini, PBT melibatkan pemprosesan teks berstruktur dan tidak berstruktur. Teks berstruktur yang digunakan dalam laman web adalah teks yang ditandai dalam format HTML dan XML yang memiliki susunan dan corak yang teratur dan mudah dianalisis. Sebaliknya, teks tidak berstruktur sukar untuk dianalisis kerana memiliki ciri yang dinamik dan bebas (Hassan et al. 2015). Teks yang tidak berstruktur dapat menghasilkan satu maklumat yang bermakna dengan bantuan PBT.

Pengecaman entiti nama (PEN) ialah satu bahagian tugas dalam PBT yang diguna untuk mengenal pasti entiti nama dan mengkelaskannya dalam kategori yang ditentukan. Entiti nama digolongkan menjadi tiga kelas iaitu Entiti Nama (ENAMEX), Numerik (NUMEX) dan Masa (TIMEX). Perbezaan ciri kelas entiti nama tersebut diklasifikasi menjadi beberapa subkelas lainnya mengikut keperluan (Aboaoga & Aziz 2013). PEN pertama kali dibahas semasa persidangan MUC-6 bagi mengenal pasti jenis entiti, seperti orang, lokasi, dan organisasi serta masa, mata wang, dan peratusan dalam teks tidak terstruktur (Abd & Mohd 2017).

LATAR BELAKANG

Penyelidikan mengenai PEN telah berkembang, satu di antara bahasa yang diguna dalam PEN ialah bahasa Inggeris. Jumlah korpus bahasa Inggeris yang banyak disedia menyebabkan ramai penyelidik menjalankan kajian PEN dalam bahasa Inggeris. Manakala korpus bahasa Melayu masih terhad berbanding sumber bahasa Inggeris (Sazali 2016). PEN bahasa Melayu tidak dapat mengguna korpus bahasa Inggeris sedia ada kerana perbezaan perkataan dan struktur morfologi antara bahasa Inggeris dan bahasa Melayu. Perbezaan perkataan ini memerlu masa lama dalam proses penerjemah agar perkataan sesuai diguna. Sulaiman et al. (2017) mengecam entiti nama bahasa Melayu mengguna dua sistem sedia ada iaitu Stanford dan Illnois yang mengguna bahasa Inggeris. Hasil kajian memperoleh nilai ketepatan yang rendah karena terdapat kesalahan bagi morfologi antara bahasa Inggeris dengan bahasa Melayu. Walau bagaimanapun, bahasa Inggeris dan bahasa Melayu memiliki kesamaan bagi penggunaan huruf besar di awal perkataan yang menunjuk kata nama khas sebagai entiti nama (Alfred et al. 2014).

Selain menggunakan korpus bahasa Inggeris, penyelidikan mengenai PEN bahasa Melayu juga telah dilaksana oleh beberapa penyelidik. Di antara pendekatan yang dilakukan bagi kajian PEN ialah pendekatan berasaskan peraturan. Peraturan dibangun oleh pakar linguistik yang mampu mengecam entiti nama dipandang daripada aspek morfologi, sintaktik atau kata kunci yang menjelaskan sifat teks (Aboaoga & Aziz 2013). Pendekatan ini mengguna corak yang dibuat secara manual kepada perkataan dalam ayat dengan mengguna satu set senarai peraturan. Alfred et al. (2014) mencadang kaedah peraturan dan kamus dalam pengecaman entiti nama bahasa Melayu. Peraturan ini menggunakan penandaan golongan kata pada bahasa Melayu untuk mengecam tiap perkataan. Apabila penandaan golongan kata untuk kata tertentu merujuk pada kata nama khas, maka peraturan tersebut diguna untuk mengenal pasti golongan perkataan tersebut. Kajian oleh Alfred et al. (2014) mengecam tiga entiti nama iaitu individu, organisasi dan lokasi. Melalui tiga fasa, kajian ini bermula dari memecahkan ayat menjadi beberapa token seperti potongan perkataan, tanda baca dan nombor. Sebagai contoh bagi mengecam entiti lokasi, Kuala Lumpur dipecah kepada dua perkataan Kuala dan Lumpur. Seterusnya menganotasi setiap token dengan mengguna peraturan berdasarkan golongan kata iaitu peraturan seperti apitan imbuhan dan merujuk kepada kamus tesaurus bahasa Melayu. Fasa akhir adalah menentukan kelas perkataan mengikut entiti nama yang telah dikenal pasti. Berdasarkan kajian ini, beberapa peraturan yang diambil kira ialah kata akhiran, kata sendi nama, dan beberapa jenis kamus. Jadual 1 menunjukkan ringkasan perbandingan kajian terdahulu yang dilakukan PEN bahasa Melayu.

JADUAL 1. Perbandingan kajian-kajian lepas

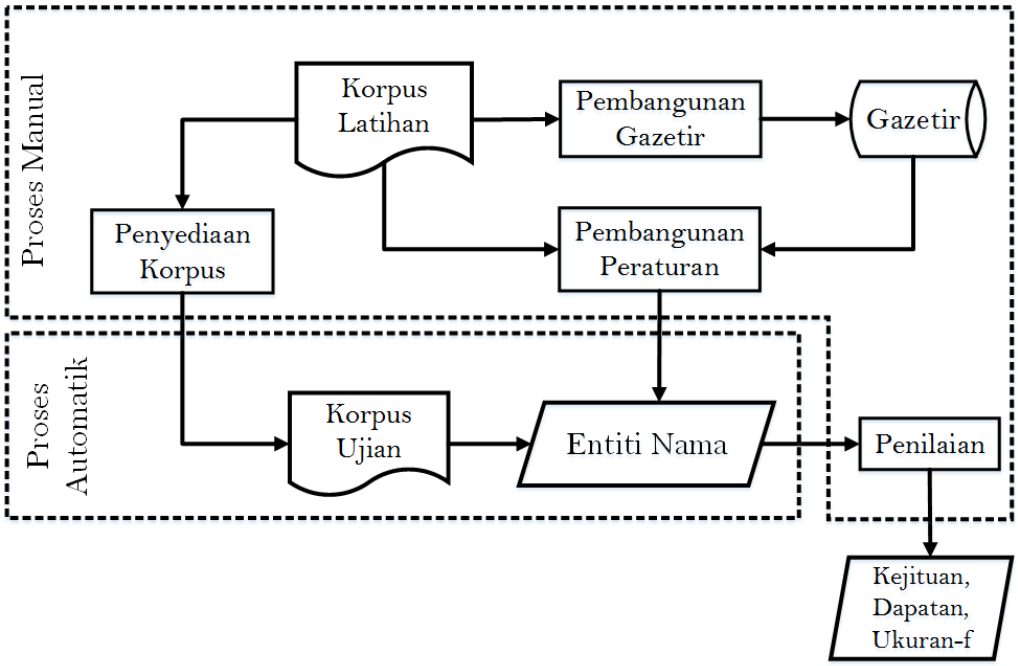
Penyelidik	Kaedah Kajian	Hasil Kajian
Alfred et al. (2014)	PEN berasaskan peraturan menggunakan kontekstual, penandaan golongan kata serta gazetir, dengan peraturan berikut: 1) PEN dikenal pasti secara berurutan dengan adanya akhiran organisasi, awalan lokasi, awalan individu. 2) Jika tidak dapat dikenalpasti maka dilakukan semakan semula pada entiti sedia ada.	Hasil ukuran-f iaitu 89.47%. <u>Kelemahan:</u> 1) Menggunakan kamus dalam pengecaman entiti mengandungi perkataan 'dan' 2) PEN bermula bagi mengenal pasti lokasi sehingga memisahkan perkataan "Medan anak Nuying"
Abd Rahman et al. (2016)	Membangun 5 peraturan dalam mengekstrak perawi hadits menggunakan *regex*.	Terdapat nama perawi yang sama dengan perawi yang mempunyai ejaan yang berbeza. Terdapat juga pelbagai bentuk nama perawi yang merujuk kepada satu orang.
Salleh et al. (2017)	Mengguna sistem pendekatan berselia dengan teknik Conditional Random Fields (CRF) yang terdiri dari 3 fasa utama 1) Pra-pemprosesan: token perkataan, frasa dan simbol. 2) Penandaan golongan kata dan melakukan pelabelan manual bagi data latihan untuk entiti nama individu, lokasi, organisasi dan sarana. 3) Membangun model menggunakan ciri perkataan seperti berawalan huruf besar, perkataan sebelum dan selepas dan golongan kata.	Hasil yang diperoleh bagi ukuran-f ialah 70%. Hal ini kerana penggunaan data latihan yang sedikit dan bergantung pula pada ciri set yang digunakan semasa membangunkan model.
Sulaiman et al. (2017).	Merujuk kepada Standford dan Illinois dalam pengecaman entiti nama individu, organisasi dan lokasi.	Standford didapati kejituan 36.55% sedangkan Illinois didapati kejituan yang lebih rendah iaitu 35.64%.
Ulanganathan et al. (2017)	Tiga langkah yang dilakukan dalam PEN iaitu 1) Pra-pemprosesan pemecahan perkataan pada data latihan, 2) Anotasi manual bagi entiti individu, lokasi, organisasi dan sarana. 3) Perkataan diproses menggunakan Linear-Chain CRF. Membandingkan hasil keputusan PEN daripada sistem Mi-Ner, PEN berasaskan peraturan dan sistem yang sedia ada iaitu Semantria.	▪ Hasil menujukkan sistem Mi-Ner diperoleh nilai kejituan 89.87% yang lebih tinggi sedikit daripada PEN berdasarkan peraturan iaitu 78.95% dan sistem Semantria 52.74%.

Kajian lepas PEN bahasa Melayu kebanyakannya menggunakan pendekatan peraturan dan pendekatan sistem berselia. Pendekatan berselia sangat bergantung dengan jumlah korpus latihan yang dapat mempengaruhi nilai ketepatan bagi PEN. Semakin banyak korpus latihan maka akan diperoleh hasil keputusan yang tinggi dan sebaliknya (Ulanganathan et al. 2017; Salleh et al. 2017). Selain itu, pendekatan berselia juga sangat bergantung pada ciri yang dipilih sebelum membangun model. Pemilihan ciri yang tidak sesuai dapat menjadi sebab kesalahan kategori entiti nama (Salleh et al. 2017).

Kaedah peraturan merupakan kaedah praktikal yang tidak memerlukan banyak data, mudah ditambah baik dan dapat dianalisis semula dengan menambahkan peraturan baru (Kumawat & Jain 2015). Meskipun kajian PEN bahasa Melayu terdahulu menghasilkan peraturan, namun beberapa peraturan yang dihasilkan masih tidak mencukupi dan tidak menyeluruh dalam mengecam entiti nama. Hal ini disebab oleh tiada piawaian peraturan yang merujuk kepada struktur bahasa Melayu (Alfred et al. 2014). Oleh itu, bagi mengatasi struktur perkataan yang kompleks, maka dibangun beberapa peraturan bagi entiti nama. Justeru, kajian ini mengambilkira penambahan jumlah peraturan dan penggunaan korpus dari pelbagai domain agar PEN dapat dilihat secara menyeluruh. Kajian ini juga menambahkan beberapa jenis entiti lainnya seperti, masa, tarikh, peratusan, kewangan dan ukuran.

METODOLOGI KAJIAN

Metodologi kajian perlu diadakan bagi memberi gambaran yang lebih jelas dan menyeluruh tentang sesuatu kaedah yang dijalankan. Proses ini terdiri daripada empat fasa iaitu penyediaan korpus, pembangunan gazetir, pembangunan peraturan dan penilaian. Rangka kerja aliran proses kajian adalah seperti yang ditunjuk dalam Rajah 1.



RAJAH 1. Rangka kerja aliran proses kajian

PENYEDIAAN KORPUS

Artikel berita yang digunakan terdiri daripada pelbagai domain berita nasional seperti kemalangan, sukan, bencana, semasa, perayaan dan politik. Sumber berita seperti Bernama, Berita Harian dan Malaysia Kini menggunakan bahasa formal dan telah disunting sesuai sebagai keperluan korpus. Pengumpulan artikel berita diekstrak secara manual ke dalam bentuk fail teks pada bahagian isi. Korpus latihan terdiri daripada 170 fail teks yang mengandungi 13987 patah perkataan dan korpus ujian sebanyak 30 fail teks dengan 7311 perkataan yang disediakan sebelum proses PEN dilaksana.

PEMBANGUNAN GAZETIR

Gazetir dibangun mengikut keperluan setiap kelas entiti nama. Antara gazetir tersebut ialah nama individu, nama jawatan, nama organisasi, dan nama syarikat. Sebagai contoh, gazetir untuk kategori individu ialah senarai nama awalan individu seperti Ahmad, Azlan, Nurul dan sebagainya dan juga digunakan nama gelaran, gelaran kehormat, gelaran diraja. Manakala gazetir yang digunakan untuk mengenal pasti organisasi ialah agensi, lembaga, persatuan, syarikat, pertubuhan, kelab, hospital, sekolah, university dan fakulti. Gazetir bagi sebahagian nama jawatan pula terdiri daripada senarai jawatan dan pekerjaan yang wujud di Malaysia. Antara senarai tersebut ialah pengurus, setiausaha dan jururawat. Manakala gazetir untuk kategori tempat ialah awalan nama tempat seperti Dewan, Taman, Jalan, dan sebagainya. Kuantiti bagi setiap gazetir mengikut kepada kewujudan sumber yang dirujuk.

PEMBANGUNAN PERATURAN

Pembangunan peraturan entiti nama dalam kajian ini memberi fokus kepada PEN yang melibatkan 9 entiti iaitu nama individu, lokasi, organisasi, jawatan, tarikh, masa, kewangan, ukuran dan peratusan. Jadual 2 menerangkan maksud daripada tiap entiti nama.

JADUAL 2. Klasifikasi entiti nama

Bil.	Entiti Nama	Penerangan	Contoh
1.	Individu	Nama orang yang dimula/diikuti dengan gelaran/pangkat.	Datuk Khairi Hassan, Prof Dr Abdul Hadi
2.	Jawatan	Nama jawatan yang diikuti nama organisasi atau nama tempat.	Ketua Polis Daerah Kota Bharu, Menteri Pembangunan Wanita, Keluarga dan Masyarakat, Ketua Penerangan PPBM
3.	Organisasi	Salah satu KNK yang berkaitan dengan organisasi, syarikat, perusahaan, pejabat, hospital, polis dan terdapat pula nama tempat yang mengikutinya.	Malaysian Industrial Development Finance (MIDF) Berhad, U-Mobile, Kementerian Perdagangan Dalam Negeri, Koperasi dan Kepenggunaan, Kolej Universiti Islam Perlis (KUIPs)
4.	Lokasi	Entiti nama daripada KNK yang menerangkan tempat	Surau Rumah Kanak-Kanak Arau, Jalan Sultan Ismail, Melaka
5.	Tarikh	Entiti yang menujukkan adanya hari, tanggal, bulan, tahun dan gabungannya.	31 Ogos 2017, September 2015, 21 Ogos
6.	Masa	Entiti nama yang berkaitan dengan waktu kejadian	jam 10.20 pagi, pukul 9.30 malam semalam.
7.	Kewangan	Entiti nama yang diawali atau diakhiri dengan tanda mata wang	AS$358 juta, 2.39 sen, RM9.95
8.	Ukuran	Entiti nama yang diikuti satuan seperti berat, tinggi, lebar	656,494 kaki persegi, 850 x 900 kaki persegi
9.	Peratusan	Entiti nama yang diakhiri dengan simbol peratus (%) atau perkataan peratus	19.7 peratus, 19.7 %

Proses pembangunan peraturan bagi pengecaman bagi setiap perkataan dalam artikel berita diproses menggunakan ungkapan nalar dengan corak tertentu dan melibatkan gazetir yang telah dibangun. Ungkapan nalar (*regex*) iaitu rentetan (*string*) teks khas bagi menentukan corak carian (Morsidi et al. 2017). Ungkapan nalar dibangun berdasarkan ciri khusus yang

diambil daripada perkataan yang cenderung sering muncul untuk menentukan penyelesaian terhadap kekangan string atau perkataan (Salleh et al. 2017). Jadual 3 menerangkan beberapa ungkapan nalar yang digunakan dalam membangun peraturan dan Jadual 4 pula menunjukkan contoh perincian peraturan entiti nama yang digunakan dalam kajian ini.

JADUAL 3. Ungkapan nalar peraturan entiti nama

Regex	Penerangan
&&	Dan
||	Atau
Eq	Sama
Ne	tidak sama
+	cocok dengan satu atau lebih karakter apapun. namun tidak cocok dengan rentetan kosong.
*	cocok dengan karakter apapun, termasuk rentetan kosong
^	Dimulakan
$	Diakhiri
=~	bernilai benar jika ungkapan nalar sesuai
!~	bernilai benar jika ungkapan nalar tidak sesuai

JADUAL 4. Contoh peraturan dan penggunaan ungkapan nalar bagi entiti nama

Jenis Entiti	Peraturan	Ungkapan nalar	Contoh penggunaan
Entiti Individu	Jika perkataan pertama (W) mengandungi perkataan gelaran daripada set A dalam gazetir dengan A= {Tengku, Nik, Haji, Tun, Tan Sri, Datuk, M., A. ...} dan perkataan seterusnya Wi+1 perkataan berawalan huruf besar maka perkataan tersebut ialah entiti nama individu. Jika Wi+1 didapati mempunyai kurungan, titik, koma atau titik koma atau perkataan daripada gazetir entiti lain, maka pengecaman untuk entiti individu selesai.	if ($arr[i] eq @gazetirIndividu && $arr[$i+1]=~/[A-Z]+[a-z]*/ && $arr[$i+2]!~ /([A-Z]+[a-z]*\.|;|\,)$/){ push @arr, “individu”;	**[individu Tuanku Syed Faizuddin Putra Jamalullail]** mengajak umat Islam seluruh dunia bersatu…
Entiti Lokasi	Jika perkataan pertama (*W)* mengandungi perkataan daripada set A dalam konteks yang merupakan kata sendi dengan A= {*di, ke, dari*} dan W_{i+1} perkataan seterusnya yang mengandungi satu atau lebih huruf besar maka perkataan tersebut ialah entiti lokasi. Jika W_{i+1} perkataan seterusnya didapati titik, koma atau titik koma dan perkataan daripada gazetir entiti lain kecuali gazetir entiti individu, maka pengecaman untuk entiti lokasi selesai.	if ($arr[i] =~/^(di\|ke\|dari)$/ && $arr[$i+1]=~ /\b[A-Z]+[a-z]*\b|^[0-9]|of/ && $arr[$i+2]!~ /['.,:;][A-Z][a-z]+\'|\.|\,|\)|\:|\;/){ push @arr, “lokasi”;	…di **[lokasi Surau Rumah Kanak-Kanak Arau]** dibawah anjuran…
Entiti Organisasi	Jika perkataan pertama (*W)* mengandungi perkataan daripada set A dalam gazetir dengan A= {*kolej, lembaga, universiti…*} dan W_{i+1} perkataan seterusnya yang mengandungi satu atau lebih huruf besar maka perkataan tersebut ialah entiti organisasi. Jika W_{i+1} didapati pembuka dan penutup kurungan maka akronim tersebut	if ($arr[i] eq @gazetirOrganisasi && $arr[$i+1]=~ / ^(\(|[A-Z]+[a-z]*) / && $arr[$i+1]!~ /['.,:;][A-Z][a-z]+\'|\.|\,|\)|\:|\;/){ push @arr, “organisasi”; }	**…[organisasi Kolej Universiti Islam Perlis (KUIPs)],**

	dikumpulkan dalam satu set entiti organisasi. Jika didapati perkataan lainnya selepas titik atau titik koma atau perkataan daripada gazetir entiti lain, maka pengecaman untuk entiti organisasi selesai.		
Entiti Jawatan	Jika perkataan pertama (*W)* mengandungi perkataan daripada set A dalam gazetir dengan A= {*Ketua, Pengerusi, Timbalan…*} dan W_{i+1} perkataan seterusnya yang mengandungi satu atau lebih huruf besar maka perkataan tersebut ialah entiti jawatan. Jika perkataan seterusnya (W_{i+1}) didapati penutup kurungan, titik dan titik koma dan perkataan daripada gazetir entiti lain kecuali entiti organisasi, maka pengecaman bagi entiti jawatan selesai.	if ($arr[i] eq @gazetirJawatan && $arr[$i+1]=~ / ^(\(|[A-Z]+[a-z]*)/ && $arr[$i+1]!~ /['.,:;][A-Z][a-z]+\'|\.|\,|\)|\:|\;/){ push @arr, "jawatan"; }	"**[jawatan Penolong Pengarah Kanan,]** Dr Xavier Jayakumar…".
Entiti Tarikh	Jika perkataan (*w)* mengandungi perkataan daripada set A dengan A= {*haribulan, ahad, isnin, selasa, ..., jan, feb, mac, ..., januari, september, nov, dis, …*} atau w_{i-1} perkataan yang mengandungi satu atau dua nombor atau w_{i+1} perkataan yang merupakan nombor yang diawali 19 atau 20 dengan dua angka selepasnya yang dikenalpasti sebagai awalan tahun maka perkataan tersebut ialah entiti tarikh.	if ($arr[i] eq @gazetirTarikh && $arr[$i-1]=~/(\d\d)\/$fh[$k+1]\/(19|20)\d\d/){ if($arr[$i-2]=~ /^(Isnin|Selasa|Rabu|Khamis|Jumaat|Sabtu|Ahad)/ && $arr[$i-1]=~/\d+/ && $arr[$i] =~ /(\W|^)$fh[$k](\W|$)/i && $arr[$i+1]=~/(\D|^)(19|20)\d\d(\D|$)/){ if($arr[$i-2]=~ /^(Isnin|Selasa|Rabu|Khamis|Jumaat|Sabtu|Ahad)/ && $arr[$i-1]=~/\d+/ && $arr[$i] =~ /(\W|^)$fh[$k](\W|$)/i){ if($arr[$i-1]=~/\d+/ && $arr[$i] =~ /(\W|^)$fh[$k](\W|$)/i && $arr[$i+1]=~/(\D|^)(19|20)\d\d(\D|$)/){ if ($arr[$i-2]=~/^(hingga\|dan)$/ && $arr[$i-3]=~/\d+/){ push @arr, "tarikh"; }	20 Mac 2014, Nov 2014 atau 20 Mac "ARAU, **[tarikh 7 Sept]**".
Entiti Masa	Jika perkataan (*w)* mengandungi perkataan daripada set B dengan B= {*pagi, malam, AM, PM, minit, saat*} dan w_{i-1} perkataan sebelumnya merupakan nombor maka perkataan tersebut ialah entiti masa.	if (clean_str($arr[$i]) =~/malam|pagi|petang|sore/ && $arr[$i-1]=~ /^\d{1,2}\.[0-5][0-9]$\|^\d+\|^(satu\|dua\|tiga\|empat\|lima\|enam\|tujuh\|lapan\|sembilan\|sepuluh)$/){	"…pada pukul **[masa 7 malam…]**", "…jam **[masa 9.30 malam**] semalam…", "antara **[masa jam 9.30 pagi hingga 10 pagi]**",.

		if (clean_str($arr[$i]) =~/^(malam|hari)/ && $arr[$i-1]=~ /^tengah/ && $arr[$i-2]=~ /^\d{1,2}\.[0-5][0-9]$|^\d+/){ push @arr, "masa"; }	
Entiti Kewangan	Jika perkataan (*w)* mengandungi perkataan daripada set A dengan A= {*ringgit, rupiah, sen, …, $, RP, VND, RM, juta, bilion…*} dan w_{i-1} perkataan yang merupakan nombor maka perkataan tersebut ialah entiti kewangan.	if ($arr[$i]=~ /\$|\bRM|RP?(?=\(.*\)|[^()]*$)\(?\d{1,3}(,?\d{3})?(\.\d\d?)?\)?$/){ push @arr, "kewangan"; }	"…**[kewangan RM187 juta]**", "kira-kira **[kewangan AS$358 juta]**…", **[kewangan 10,000 Dong Vietnam (VND)]** setiap satu untuk pertimbangan pembelian **[kewangan 121.62 bilion VND (RM21.28 juta)]**".
Entiti Ukuran	Jika perkataan (*w)* mengandungi perkataan daripada set A dengan A= {*liter, meter, kaki, …, cm, kg, km, …*} dan w_{i-1} perkataan yang merupakan nombor maka perkataan tersebut ialah entiti kewangan.	if ($arr[i] eq @gazetirUkuran && $arr[$i-1]=~/^\d+/ && $arr[$i-2]=~/\d+/ && $arr[$i-3]!~ /['.,:;][A-Z][a-z]+\'|\.|\,|\)|\:|\;/){ push @arr, "Ukuran"; }	"…kawasan lantai kasar keseluruhan sebanyak **[ukuran 656,494 kaki persegi]**", "…kawasan lantai **[ukuran 850 x 900 kaki persegi]**" atau "Unit Rumawip yang berukuran **[ukuran 79 meter]**…",
Entiti Peratusan	Jika perkataan (*w)* mengandungi perkataan daripada set A dengan A= {*peratus, %*} dan w_{i-1} perkataan	if (clean_str($arr[$i]) =~/^\d+(\.\d+)?$\|^(satu\|dua\|tiga\|empat\|lima\|enam\|tujuh\|lapan\|sembilan\|sepuluh)$/ && $arr[$i+1] =~/peratus/){ if ($arr[$i] =~/^\d+(\.\d+)?%$/){ push @arr, "peratusan"; }	"…projek pembangunan semula **[peratusan 30 peratus]**

PENILAIAN

Pengujian dan penilaian dilakukan terhadap hasil dapatan kajian yang diperoleh dari fasa sebelumnya. Setiap peraturan pengecaman entiti nama yang dibangunkan diuji dengan menggunakan kedua-dua jenis korpus. Korpus latihan digunakan semasa pembangunan peraturan dan hasilnya diuji dengan menggunakan korpus ujian bagi menilai tahap ketepatan PEN. Keputusan dinilai dari segi kejituan, dapatan dan ukuran-f yang dapat dirumuskan dengan menggunakan formula (1), (2) dan (3) (Alfred et al. 2014).

$$Kejituan = \frac{tepat + (0.5 * separa\ tepat)}{jumlah\ keseluruhan\ PEN\ oleh\ sistem}, \quad (1)$$

$$Dapatan = \frac{tepat + (0.5 * separa\ tepat)}{jumlah\ keseluruhan\ PEN\ secara\ manual}, \quad (2)$$

$$Ukuran - F = \frac{(kejituan * dapatan)}{0.5 * (kejituan + dapatan)}. \quad (3)$$

HASIL ANALISIS DAN PERBINCANGAN

Berdasarkan pengujian yang dijalankan terhadap data korpus ujian diperoleh keputusan dari aspek kejituan, dapatan dan ukuran-F. Hasil keputusan yang diperoleh daripada pengujian seperti yang diringkas pada Jadual 5.

JADUAL 5. Keputusan di antara ketiga penilaian

Jenis entiti	Dapatan	Kejituan	Ukuran-f
Individu	88.30%	86.90%	87.60%
Organisasi	97.42%	93.56%	95.45%
Jawatan	93.66%	92.36%	93%
Lokasi	93.08%	93.85%	93.46%
Tarikh	97.77%	97.77%	97.77%
Masa	92.85%	92.85%	92.85%
Peratusan	98.21%	98.21%	98.21%
Ukuran	92.85%	92.85%	92.85%
Kewangan	94.44%	88.54%	91.39%
Purata	94.29%	92.99%	93.62%

Keputusan yang diperoleh, menunjukkan nilai kejituan bagi dapatan, kejituan dan ukuran-f adalah sebanyak 94.29%, 92.99% dan 93.62%. PEN bagi entiti individu diperoleh nilai kejituan iaitu 86.90%, dapatan 88.30% dan ukuran-f 87.60%. Kesalahan berlaku sewaktu perulangan entiti nama yang berbeza huruf dari entiti nama individu sebelumnya. Sebagai contoh, "Datuk Shahruddin Khalid mengatakan… hal ini dipengaruhi oleh Shahrudin". Perkataan "Shahruddin" merupakan sebahagian daripada perkataan "Datuk Shahruddin Khalid", akan tetapi terjadi kesalahan penulisan menjadi "Shahrudin". Kedua perkataan ini dianggap berbeza dan tidak dapat dikenalpasti oleh peraturan yang sedia ada. Selain itu, penggunaan entiti nama yang tidak wujud dalam senarai gazetir juga dapat menurunkan nilai ketepatan. Sebagai contoh, "Timbalan Dekan Universiti Teknikal Malaysia Melaka, Fifiyana Asmida bersama rakannya mengadakan" perkataan "Fifiyana Asmida" tidak dapat dicam kerana perkataan tersebut tidak wujud dalam senarai gazetir.

Seterusnya, bagi entiti organisasi nilai kejituan, dapatan dan ukuran-f ialah 93.56%, 97.42% dan 95.45%. Pada pengujian ini telah berlaku rujukan silang antara entiti nama individu dan lokasi. Sebagai contoh "Hospital Tun Aminah" dan "Balai Polis Negeri Sembilan" mengandungi nama individu atau lokasi dalam entiti organisasi. Kesalahan pengecaman yang berlaku kerana nama organisasi mengguna perkataan bahasa Inggeris yang tidak wujud dalam senarai gazetir, sebagai contoh "Fuji Electrics".

Bagi entiti jawatan didapati nilai kejituan, dapatan dan ukuran-f ialah 92.36%, 93.66% dan 93%. Hal ini ialah kerana adanya peraturan bagi penggunaan entiti nama lain pada entiti jawatan, seperti entiti organisasi dan lokasi. Sebagai contoh, "Ketua Polis Daerah Kemaman" atau yang mengandungi entiti nama organisasi "Pengarah Jabatan Kemajuan Orang Asli (JAKOA) Negeri Kelantan". Adapun terjadi kesalahan kerana pada jawatan mengandungi nama individu seperti "Presiden Pertubuhan India Koridor Utara (NIC), G Rajasegar". Perkataan "G Rajasegar" merupakan entiti nama, ini berlaku kerana pada peraturan yang telah dibuat untuk entiti jawatan diikuti simbol koma dan selepas koma (,) huruf besar sehingga perkataan "G Rajasegar" termasuk entiti jawatan. Selain itu, pengecaman bagi entiti jawatan juga akan selesai jika terdapat huruf kecil atau mengandungi entiti lainnya selain organisasi dan lokasi. Perkataan "G Rajasegar" tidak wujud dalam senarai gazetir individu sehingga perkataan tersebut dicam sebagai entiti jawatan.

Bagi entiti lokasi, nilai kejituan, dapatan dan ukuran-f yang diperoleh iaitu 93.85%, 93.09% dan 93.47%. Hal ini kerana pada lokasi dapat juga mengandungi entiti nama lainnya seperti nama individu. Sebagai contoh, "Jalan Sultan Ismail" yang dapat dijadikan satu entiti dan tidak dipisahkan. Namun begitu, kesalahan pengecaman dapat berlaku manakala jika entiti nama tidak terdapat dalam gazetir, misalnya "KLIA2 Sepang".

Bagi entiti tarikh didapati nilai kejituan, dapatan dan ukuran-f ialah 97.77%, 97.77% dan 97.77%. Kesemua entiti tarikh dapat dikenalpasti dengan tepat dan hanya satu yang tidak dapat dikenalpasti seperti "1990-an". Hal ini kerana pada peraturan yang dibuat tidak mengandungi tahun dengan akhiran "–an". Seterusnya, bagi entiti masa pula diperoleh nilai kejituan, dapatan dan ukuran-f yang sama iaitu 92.85%. Manakala bagi entiti peratusan diperoleh nilai 98.21% bagi kejituan, dapatan dan ukuran-f. Entiti ukuran pula didapati nilai kejituan, dapatan dan ukuran-f yang sama iaitu 92.85%. Kesalahan dalam pengecaman ukuran iaitu perkataan yang mengandungi satuan "km/s" kerana tidak adanya perkataan tersebut dalam senarai gazetir ukuran.

Beberapa entiti seperti tarikh, masa, peratusan dan ukuran diperoleh nilai dapatan, kejituan dan ukuran-f yang sama. Hal ini kerana jumlah keseluruhan PEN oleh sistem dan jumlah keseluruhan PEN secara manual adalah sama banyak sehingga diperoleh nilai peratusan yang sama. Seterusnya, PEN berasaskan peraturan yang telah dibangunkan oleh Alfred et al. (2014) dijadikan sebagai perbandingan penilaian bagi menilai peraturan PEN yang telah dibangunkan. Perbandingan penilaian ini menggunakan korpus yang diperoleh daripada Alfred et al. (2014). Korpus ini mengandungi 7325 perkataan yang diuji hanya pada tiga jenis entiti iaitu entiti individu, lokasi dan organisasi. Jadual 6 menunjukkan hasil keputusan menggunakan korpus daripada Alfred et al. (2014) dan Jadual 7 menunjukkan keputusan perbandingan yang diperoleh dalam kajian ini.

JADUAL 6. Hasil keputusan korpus Alfred et al. (2014)

Jenis Entiti	Dapatan	Kejituan	Ukuran-f
Individu	93.91%	82.73%	87.97%
Organisasi	90.90%	93.75%	92.30%
Lokasi	91.58%	94.23%	92.89%
Nilai purata	92.13%	90.23%	91.05%

JADUAL 7. Perbandingan keputusan antara dua PEN

Jenis PEN	Dapatan	Kejituan	Ukuran-f
PEN kajian ini	92.13%	90.23%	91.05%
PEN Alfred et al. (2014)	94.44%	85%	89.47%

Dari segi ukuran-f pula, PEN dalam kajian ini sedikit lebih tinggi iaitu 91.05% berbanding PEN kajian Alfred et al. (2014). Hal ini kerana ukuran-f dipengaruhi oleh nilai kejituan dan dapatan (Alfred et al. 2013). Kejituan yang diperoleh kajian ini sebanyak 90.23% dengan perbezaan yang ketara daripada Alfred et al. (2014) iaitu 5.23%. Kajian ini telah membangunkan peraturan PEN bagi penggunaan perkataan 'dan' atau simbol &, seperti "Jabatan Keselamatan dan Kesihatan Pekerjaan" yang merupakan satu entiti nama organisasi. Manakala pada kajian Alfred et al. (2014) pengecaman entiti nama yang mengandungi perkataan 'dan' tidak dapat diselesaikan sehingga diperlukan kamus khas yang berisi senarai entiti yang mengandungi perkataan 'dan', akan tetapi kamus khas yang dibangunkan tidak mampu bagi menghadapi perubahaan pada entiti nama.

KESIMPULAN

Kajian ini membangunkan peraturan baru bagi PEN bahasa Melayu dan membandingkan PEN yang dibangunkan dalam kajian ini dengan kajian lepas bahasa Melayu berasaskan peraturan. Termasuk kehadiran entiti nama bersama perkataan 'dan' dan simbol, pengulangan entiti nama pendek, penggunaan entiti nama yang mengandungi entiti lain dan percampuran entiti nama yang berbeza. Hasil uji kaji menunjukkan keputusan dapatan, kejituan dan ukuran-f PEN yang dibangunkan dalam kajian ini adalah lebih baik dengan nilai kejituan 92.13%, 90.23% dan 91.05% berbanding nilai kejituan kajian lepas iaitu sebanyak 94.44%, 85% dan 89.47%. Ini

menunjukkan bahawa jumlah peraturan yang digunakan mempengaruhi nilai ketepatan PEN. Oleh itu, dengan mengubah suai peraturan dalam kajian ini dan menambahkan peraturan baru diharapkan dapat membantu penyelidik lain dalam melaksanakan PEN bagi korpus bahasa Melayu untuk menghasilkan nilai ketepatan yang tinggi.

RUJUKAN

Abd, M. T. & Mohd, M. 2017. Extended Distributed Prototypical for Biomedical Named Entity Recognition 6(2): 1–11.

Abd Rahman, N., Alias, N., Ismail, N. K., Bin Mohamed Nor, Z. & Alias, M. N. B. 2016. An identification of authentic narrator's name features in Malay hadith texts. *ICOS 2015 - 2015 IEEE Conference on Open Systems* (August): 79–84. doi:10.1109/ICOS.2015.7377282

Aboaoga, M. & Aziz, M. J. A. 2013. Arabic person names recognition by using a rule based approach. *Journal of Computer Science* 9(7): 922–927. doi:10.3844/jcssp.2013.922.927

Alfred, R., Leong, L. C., On, C. K. & Anthony, P. 2014. Malay Named Entity Recognition Based on Rule-Based Approach 4(3). doi:10.7763/IJMLC.2014.V4.428

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

Halid, N. A. & Omar, N. 2017. MALAY PART OF SPEECH TAGGING USING RULED-BASED APPROACH. Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik 6(2): 91–107.2[[

Hassan, M., Nazlia, O. & Mohd Juzaiddin, A. A. 2015. Malay Part of Speech Tagger : A Comparative Study on Tagging Tools. Asia-Pacific Journal of Information Technology and Multimedia 4(1): 11–23. doi:10.17576/apjitm-2015-0401-02

Jiang, R., Star, A. & Banchs, R. E. 2016. Evaluating and Combining Named Entity Recognition Systems 21–27.

Jones, K. S. 2001. Natural Language Processing : a Historical Review. *Natural Language Processing : a Historical Review*.

Kumawat, D. & Jain, V. 2015. POS Tagging Approaches : A Comparison 118(6): 32–38.

Morsidi, F., Sulaiman, S. & Abdul, R. 2017. Feature Extraction using Regular Expression in Detecting Proper Noun for Malay News Articles based on KNN Algorithm (June), 0–23. doi:10.4314/jfas.v9i5s.16

Salleh, M. S., Asmai, S. A., Basiron, H. & Ahmad, S. 2017. A Malay Named Entity Recognition Using Conditional Random Fields. *International Conference on Information and Communication Technology (ICoICT) A* 0(c).

Sazali, S. S. B. 2016. Information Extraction : Evaluating Ner From Classical Malay Documents. *International Conference on Information Retrieval and Knowledge Management Information* 48–53.

Suhaimi Ab Rahman, Nazlia Omar, M. J. A. A. 2014. Extraction of Compound Nouns in Malay Noun Phrases Using a Noun Phrase Frame Structure 3(June 2014): 23–32.

Sulaiman, S., Wahid, R. A., Sarkawi, S. & Omar, N. 2017. Using SYUR4tanford NER and Illinois NER to Detect Malay Named Entity Recognition 9(2): 2–5. doi:10.7763/IJCTE.2017.V9.1128.

Ulanganathan, T., Ebrahimi, A., Chu, B., Xian, M., Bouzekri, K., Mahmud, R. & Hoe, O. H. 2017. Benchmarking Mi-NER: Malay Entity Recognition Engine Department of Artificial Intelligence University of Malaya (c): 52–58.

Yong, S., Ranaivo-malançon, B. & Wee, A. Y. 2011. NERSIL : the Named-Entity Recognition System for Iban Language 549–558.

Ulfa Nadia
Nazlia Omar
Fakulti Teknologi dan Sains Maklumat (FTSM),
UKM, Bangi, 43600, Selangor, Malaysia.
ulfa.nadia07@gmail.com, nazlia@ukm.edu.my

Received: 15 February 2019
Accepted: 30 May 2019
Published: 27 June 2019