

Penentuan Fitur bagi Pengekstrakan Tajuk Berita Akhbar Bahasa Melayu

Shahrul Azman Mohd Noah

samn.ukm@gmail.com

Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia

Nazlena Mohamad Ali

nazlena.ali@ukm.edu.my

Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia

Mohd Sabri Hasan

sabri@kabinet.gov.my

Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia

ABSTRAK

Ringkasan tajuk berita (*headline*) adalah salah satu teknik ringkasan teks automatik yang boleh mengurangkan masalah kebanjiran maklumat dalam sistem capaian. Teknik ini berupaya mengurangkan beban kognitif pengguna semasa meneliti dan memilih dokumen relevan dalam kuantiti yang besar. Keupayaan teknik ini dipengaruhi oleh ciri-ciri sistem bahasa tabii yang mewakili maklumat dalam dokumen. Kajian ini membincangkan proses dalam penentuan ciri-ciri sistem bahasa Melayu pada dokumen genre berita. Metodologi kajian dimulai dengan analisis ke atas korpus dokumen berita bahasa Melayu. Korpus ini mengandungi 140 dokumen berita teras yang dipilih daripada dua pangkalan data berita arus perdana di Malaysia iaitu Berita Harian dan Utusan Malaysia. Kriteria pemilihan adalah kategori berita teras, bersaiz 50 hingga 250 perkataan, dengan tahun penerbitan dari 2007 hingga 2012 dan genre berita adalah ekonomi, jenayah, pendidikan dan sukan. Tiga pakar linguistik bahasa Melayu menghasilkan satu ringkasan tajuk berita bagi setiap dokumen berita secara manual. Ketiga-tiga pakar linguistik ini perlu mematuhi tiga syarat iaitu ringkasan dilakukan secara pengekstrakan, teknik pemilihan perkataan secara *select-word-inorder* dan perubahan morfologi perkataan. Hasil eksperimen menunjukkan tiga fitur telah dikenal pasti iaitu, pertama: dua ayat pertama adalah calon sesuai ayat terpenting, kedua: ayat mengandungi takrifan akronim berpotensi sebagai ayat terpenting dan ketiga: saiz ringkasan tajuk berita ideal adalah enam perkataan. Pertimbangan fitur ini membolehkan ringkasan tajuk berita dijana secara automatik yang lebih mirip seperti dilakukan oleh manusia.

Kata kunci: isi utama; pemprosesan bahasa tabii; berita Bahasa Melayu; ringkasan teks; korpus bahasa melayu

Determining Features of News Headline in Malay News Document

ABSTRACT

Headline summarization is one of the automated text summarization techniques that can reduce the problem of information overload in the retrieval system and reduce the user's cognitive burden while searching and selecting relevant documents in large quantities. This study discusses the process on the determination of Malay language system features in the news genre document. Methodology starts with analysis the corpus of Malay news

documents. The corpus contains 140 core news items which were selected from the two mainstream news databases in Malaysia which are Berita Harian and Utusan Malaysia. The selection news criteria are from core news categories, sized 50 to 250 words, the years of publication from 2007 to 2012 and news genres from economic, crime, education and sports. Three linguistic experts in Malay produced a headline summary for each news document manually. The experts need to comply with three conditions which are summary extraction, select-word-inorder word selection techniques and word morphological changes. The experimental results show that three characteristics have been identified, first: the first two sentences are the important sentences, second: the verse that contains the potential acronym definitions is chosen as the most important sentence and third: the size of the summary of the ideal headline is six words. The consideration of this feature allows a summary of the headline that can be generated automatically, just like the process done by human.

Keywords: headline; Natural Language Processing; malay news; text summarization; Malay corpus

PENGENALAN

Teknologi maklumat dan rangkaian internet hari ini membolehkan maklumat dicipta, diproses, disalin dan disebar dengan mudah dan pantas dalam pelbagai versi bahasa tabii. Kepelbagaian versi bahasa tabii telah menyebabkan masalah kebanjiran maklumat dalam sistem capaian menjadi lebih kritikal. Masalah ini berpunca daripada perbezaan sistem bahasa tabii untuk mengungkapkan sesuatu makna. Walau bagaimanapun, kepesatan teknologi juga telah banyak membantu mengembangkan kajian dalam bidang linguistik. Kajian yang melibatkan teknologi telah menjadikan penyelenggaraan data lebih meluas serta terancang, sistematik dan berkesan (Hishamudin & Norsimah, 2011; Nor Hashimah & Ahmad, 2009).

Teknik ringkasan tajuk berita merupakan salah satu teknik ringkasan teks automatik yang berupaya mengurangkan masalah tersebut. Keupayaan teknik ini menjana dokumen versi ringkas yang mampu mengurangkan beban kognitif dan masa pengguna membaca dan meneliti setiap dokumen dalam sistem capaian. Ini adalah kerana dokumen versi ringkas memfokus kepada kandungan maklumat penting sahaja dan pengulangan maklumat yang sama diminimumkan.

Menurut (Kaikhah, 2004; Alireza & Moses, 2013; Gunawan, Pasaribu, Rahmat & Budiarto, 2017), teknik ringkasan tajuk berita berpotensi mengurangkan masalah kebanjiran maklumat kerana teknik ini berupaya menguruskan dokumen relevan dengan keperluan maklumat pengguna dalam kuantiti yang besar dengan mudah dan cepat. Teknik ini juga berupaya mengurangkan beban kognitif pengguna ketika meneliti dan memilih dokumen relevan dalam kuantiti yang besar.

Format ringkasan tajuk berita adalah satu ayat tunggal yang mewakili idea utama penulisan dokumen, di mana teknik ini digunakan dengan meluas dalam penjanaan tajuk bagi dokumen genre berita secara automatik. Walau bagaimanapun, secara praktisnya dalam dunia kewartawan ringkasan tajuk berita perlu diubahsuai supaya memenuhi keutamaan yang telah ditetapkan oleh agensi akhbar, fokus meja berita dan jangkauan kelompok pembaca (Saiful Nujaimi, 2009).

Kaedah pengekstrakan dokumen tunggal untuk menjana ringkasan tajuk berita melibatkan tiga proses iaitu pemilihan ayat terpenting, pemilihan perkataan penting dan realisasi ayat (Shen, 2009). Kekompleksan kaedah ini bergantung sepenuhnya kepada sistem bahasa tabii yang mewakili maklumat dalam dokumen. Pemilihan ayat terpenting adalah proses untuk menentukan ayat yang mengandungi idea utama penulisan. Pemilihan perkataan penting pula adalah proses menentukan perkataan dalam ayat terpenting yang perlu wujud

dalam ringkasan tajuk berita. Realisasi ayat adalah proses akhir yang melibatkan penyesuaian morfologi dan penyepaduan perkataan yang telah dipilih untuk membentuk satu kesatuan ayat yang sempurna.

Kajian literatur telah dijalankan dan didapati hanya satu kajian penentuan fitur sistem bahasa tabii dalam konteks teknik ringkasan headline telah dilakukan oleh (Zajic, Dorr & Schwartz, 2005). Kajian tersebut dilakukan pada dokumen genre berita bahasa Inggeris. Hasil kajian tersebut telah digunakan dalam kebanyakan teknik ringkasan headline pada hari ini. Masalah timbul apabila teknik sedia ada digunakan pada dokumen genre berita selain dari Bahasa Inggeris. Ini adalah disebabkan oleh perbezaan sistem bahasa tabii dalam mengungkapkan sesuatu makna dalam ayat.

Oleh itu, objektif kajian ini adalah untuk menentukan fitur sistem bahasa Melayu yang perlu dipertimbangkan dalam teknik ringkasan tajuk berita, khusus kepada dokumen genre berita bahasa Melayu. Kaedah yang digunakan dalam eksperimen ini diadaptasi daripada eksperimen yang telah dilakukan oleh (Zajic, Dorr & Schwartz, 2005). Hasilnya merupakan fitur sistem bahasa Melayu yang perlu dipertimbangkan dalam teknik ringkasan tajuk berita untuk menghasilkan ringkasan tajuk berita yang lebih mirip seperti dilakukan oleh manusia.

KAJIAN LEPAS

Kajian terawal ringkasan teks automatik telah dimulakan oleh Luhn (1958). Beliau menggunakan kaedah statistik sepenuhnya untuk menghasilkan dokumen versi ringkas. Kajian tersebut kurang mendapat perhatian dalam bidang sains komputer sehingga perkembangan pesat internet pada tahun 1990an yang menyebabkan masalah kebanjiran maklumat dalam capaian maklumat sedia ada. Faktor utama masalah ini adalah kesediaan teknologi membenarkan maklumat dicipta dalam pelbagai versi bahasa tabii. Oleh yang demikian, kajian ini mula dikembangkan dalam capaian maklumat, pemprosesan bahasa tabii dan kecerdasan buatan (Tengku Mohd, 2007; Steinberger & Jezek, 2009; Shahrul Azman, Nazlia & Amru Yusrin, 2015).

Kajian teknik ringkasan tajuk berita pada awalnya ditumpukan kepada dokumen genre berita bahasa Inggeris (Zajic, Dorr & Schwartz, 2002; Zhou & Hovy 2003; Dorr, Zajic & Schwartz, 2003; Kaikhah 2004). Masalah mula dihadapi apabila teknik tersebut digunakan pada dokumen genre berita yang bukan dalam Bahasa Inggeris. Dorr dan Zajic (2003) dan Dausidia dan Lapalme (2004) adalah perintis idea untuk menggabungkan teknik ringkasan tajuk berita sedia ada dengan mesin terjemah. Berdasarkan idea itu, dokumen berita diterjemahkan oleh mesin terjemah kepada teks bahasa Inggeris dan seterusnya dokumen versi ringkas dijana oleh teknik ringkasan tajuk berita sedia ada. Namun demikian, oleh kerana faktor sistem bahasa tabii, hasilnya tidak dapat diterjemahkan semula kepada bahasa asal.

Idea penggabungan ini dibuat dengan harapan menghasilkan teknik ringkasan teks automatik pelbagai bahasa tabii tanpa penglibatan manusia. Faktor keunikan bahasa tabii menjadi penghalang utama bagi merealisasikan idea penggabungan tersebut. Terdapat tiga kelemahan idea gabungan ini iaitu kesukaran dalam menentukan ayat penting bagi dokumen yang telah diterjemah (Dorr & Zajic, 2003), ketepatan penentuan ayat terpenting adalah berdasarkan ketepatan mesin terjemah (Dausidia & Lapalme, 2004) dan masalah kohesi ayat yang telah diterjemah menyebabkan makna sebenar tidak dapat dipelihara (Shamsfard et al. 2009). Oleh itu, para penyelidik mula menjalankan kajian teknik ringkasan tajuk berita mengikut konteks bahasa tabii yang mewakili maklumat dalam dokumen tanpa melibatkan mesin terjemah. Alotaiby, Foda dan Alkharashi (2011) dan Azmi dan Al-Thanyyan (2012)

telah membuktikan teknik ringkasan teks automatik dengan mempertimbangkan ciri bahasa Arab tanpa penglibatan mesin terjemah memberikan hasil yang lebih baik.

Kajian literatur menunjukkan, hanya Norshuhani dan Arina (2011) telah menjalankan kajian teknik ringkasan teks automatik secara generik khusus kepada dokumen versi bahasa Melayu. Namun, teknik yang digunakan masih menggabungkan mesin terjemah tanpa mengambil kira ciri sistem bahasa Melayu yang mewakili maklumat dalam dokumen. Disebabkan faktor sistem bahasa tabii, hasilnya tidak dapat diterjemah semula kepada teks bahasa Melayu. Manakala kajian oleh Suraya, Siti Kharijah dan Hoon (2018) mengetengahkan pendekatan generik perwakilan teks (ketidak bergantung pada bahasa) yang boleh digunakan dalam peringkasan teks, tetapi hasil yang diperoleh tidak konsisten untuk Bahasa Melayu.

Menurut Shen (2009), dan Alicja, Vania dan Katja (2017), kajian terkini ringkasan tajuk berita lebih ditumpukan kepada pengekstrakan maklumat yang menggabungkan teknik ringkasan tajuk berita, pemprosesan bahasa tabii dan pembelajaran mesin. Terdapat dua kaedah pengekstrakan maklumat, iaitu kaedah pengekstrakan maklumat tanpa penyeliaan dan kaedah pengekstrakan maklumat dengan penyeliaan.

Kaedah pengekstrakan dengan penyeliaan adalah kaedah yang memerlukan set data latihan untuk menentukan ayat penting dalam teks. Model akan dibina menggunakan fitur yang telah dikenal pasti bagi membolehkan ramalan pemilihan ayat penting dibuat. Model tersebut juga harus berupaya belajar daripada set data latihan mengenai bagaimana meramalkan ayat penting dalam teks. Seterusnya, dengan menggunakan model tersebut ayat penting dalam dokumen akan diramalkan dan dokumen versi ringkas akan dijanakan. Kaedah pengekstrakan tanpa penyeliaan adalah kaedah yang tidak memerlukan sebarang set data latihan. Kaedah ini akan menggunakan gabungan bukti yang wujud pada aras permukaan teks untuk pemilihan ayat penting dalam teks. Ia terbahagi kepada tiga iaitu kaedah pemilihan ayat penting berdasarkan kedudukan ayat, frekuensi perkataan atau frasa dan pemusatan wacana.

Oleh itu, kaedah terkini ringkasan teks automatik khususnya ringkasan tajuk berita memerlukan pemahaman yang jelas terhadap ciri sistem bahasa tabii yang mewakili maklumat dalam dokumen. Pemahaman ciri ini melibatkan morfologi pembentukan perkataan, sintaksis pembentukan struktur ayat dan rumus sintaksis yang perlu dipatuhi dalam menerbitkan ayat yang sempurna. Pemahaman ciri sistem bahasa tabii, teknik pemprosesan bahasa tabii dan teknik pembelajaran mesin membolehkan teknik ringkasan tajuk berita yang pintar dapat dibangunkan.

STRUKTUR BERITA

Definisi dokumen berita adalah luas dan pelbagai kerana dunia kewartawanan sentiasa hidup dan berkembang seiring dengan perubahan masa (Saiful Nujaimi, 2009). Namun, terdapat empat definisi yang menjadi teras kepada definisi dokumen berita iaitu (i) sesuatu yang baharu berlaku atau terjadi, (ii) laporan berhubung sesuatu acara secara kolektif, (iii) sesuatu perkhabaran yang baharu berhubung dengan apa sahaja, dan (iv) surat khabar itu sendiri adalah dokumen berita (Julian & Stanley, 1965).

Dokumen berita adalah dokumen yang paling kerap diterbitkan sama ada secara cetakan fizikal atau berbentuk format digital yang dipaparkan di portal rasmi agensi penerbitan berita. Menurut Brooks (1996), setiap dokumen berita perlu memiliki enam sifat iaitu (i) penting, (ii) lengkap, (iii) benar, (iv) cermat, (v) cepat, dan (vi) objektif. Enam sifat dokumen berita ini menjadikan dokumen berita sebagai dokumen berstruktur dan mempunyai format tertentu. Sifat penting dokumen berita bermaksud dokumen berita yang dihasilkan perlu menepati keutamaan yang ditetapkan oleh tiga pihak iaitu (i) agensi penerbitan akhbar, (ii) meja berita, dan (iii) kelompok pembaca (Saiful Nujaimi, 2009). Keutamaan yang

ditetapkan oleh agensi penerbitan akhbar merujuk kepada keutamaan agenda penerbitan akhbar yang mempertimbangkan sensitiviti agensi dan rakan kongsi agensi. Keutamaan kepada pihak meja berita pula merujuk kepada keutamaan perspektif meja berita yang diwakilinya. Manakala keutamaan kepada kelompok pembaca adalah keutamaan untuk menepati jangkauan kelompok pembaca. Jika jangkauan ini tidak dipenuhi, kemungkinan pembaca memilih akhbar lain.

Menurut Md Salleh (1985), sifat lengkap dokumen berita terdiri daripada enam elemen yang perlu wujud dalam setiap kandungan berita iaitu:

<i>What</i>	Apakah yang berlaku atau akan berlaku ?
<i>Where</i>	Di mana ia berlaku atau akan berlaku ?
<i>When</i>	Bila ia berlaku atau akan berlaku ?
<i>Who</i>	Siapa sumber yang terlibat atau siapa pelaku ?
<i>Why</i>	Mengapa ia berlaku atau akan berlaku ?
<i>How</i>	Bagaimana ia berlaku atau akan berlaku ?

Walau bagaimanapun, enam elemen ini tidak semestinya diletakkan dalam satu ayat serentak.

Saiful Nujaimi (2009) menyatakan bahawa sifat pantas dokumen berita merujuk kepada keadaan di mana dokumen berita yang diterbitkan perlu disebar dengan pantas. Kelambatan penerbitan dan penyebaran dokumen berita menyebabkan definisi dan tafsiran berita menjadi salah. Sifat objektif dokumen berita merujuk kepada penerbitan dokumen berita yang mempertimbangkan dua aspek iaitu (i) kredibiliti dan reputasi, dan (ii) integriti dan akauntabiliti. Aspek kredibiliti dan reputasi menggambarkan sumber dokumen berita perlu memiliki kredibiliti yang tinggi bagi meningkatkan reputasi berita, wartawan dan seterusnya agensi penerbitan berita. Aspek integriti dan akauntabiliti pula menggambarkan wartawan sebagai pelapor yang berintegriti di mana wartawan perlu sedar bahawa mereka dipertanggungjawabkan atas hasil kerja mereka hari ini dan pada masa akan datang sekalipun wartawan tersebut telah bertukar kerja atau telah meninggal dunia.

METODOLOGI

Korpus dokumen berita Pertubuhan Berita Nasional Malaysia (BERNAMA) merupakan korpus dokumen berita Bahasa Melayu terbesar di Malaysia. Korpus ini dapat diakses menggunakan aplikasi yang dikenali sebagai *Berita Bernama Library dan Infolink Service* (BLIS). BLIS merupakan salah satu daripada perkhidmatan kandungan yang ditawarkan oleh pihak BERNAMA. Kebanyakan akhbar dan media elektronik di Malaysia dan luar negara merupakan pelanggan utama BERNAMA. Dalam kajian ini set data yang dibangunkan terdiri daripada 140 dokumen berita yang mewakili genre berita ekonomi, jenayah, pendidikan dan sukan. Dokumen berita ini dipilih dari arkib berita BERNAMA melalui aplikasi BLIS.

Korpus yang digunakan dalam eksperimen ini terdiri daripada 140 dokumen genre berita bahasa Melayu. Dokumen berita ini dipilih daripada dua agensi pembekal berita arus perdana di Malaysia iaitu Berita Harian dan Utusan Malaysia. Terdapat empat kriteria yang telah ditetapkan dalam pemilihan dokumen genre berita iaitu kategori berita, saiz berita, tahun penerbitan dan genre berita. Kategori berita teras dipilih kerana berita teras merupakan berita utama dan 80% berita kategori ini diterbitkan untuk setiap penerbitan. Kesemua berita teras yang dipilih adalah bersaiz di antara 50 hingga 250 perkataan, dan diterbitkan pada tahun 2007 hingga 2012. Tiga puluh lima berita teras dipilih bagi setiap genre berita ekonomi, jenayah, pendidikan dan sukan. Agensi penerbitan akhbar di Malaysia lazimnya mempunyai 10 meja berita dan dibahagikan kepada lima kategori menurut Saiful Nujaimi

(2009). Daripada lima kategori tersebut, untuk dokumen berita teras majoritinya berada dalam kategori dua dan tiga. Meja berita kategori dua terdiri daripada berita genre jenayah dan sukan. Manakala meja berita kategori tiga pula terdiri daripada berita genre pendidikan, hiburan, ekonomi dan kota. Bilangan dokumen berita mengikut kategori dan meja berita yang dipilih dipaparkan di dalam Jadual 1.

JADUAL 1. Bilangan dokumen berita berdasarkan kategori dan meja berita

Kategori	Meja Berita	Bilangan Dokumen
Kategori 2	Meja Jenayah	35
Kategori 2	Meja Sukan	35
Kategori 3	Meja Ekonomi	35
Kategori 3	Meja Pendidikan	35

Untuk menghasilkan ringkasan rujukan tunggal bagi setiap dokumen berita, tiga pakar linguistik Bahasa Melayu dipilih berdasarkan kredibiliti pakar daripada aspek latihan secara formal yang diterima. Tiga orang graduan Sarjana Sastera (Kajian Bahasa Melayu) dari Pusat Pengajian Bahasa, Kesusasteraan dan Kebudayaan Melayu, Fakulti Sains Sosial dan Kemanusiaan (FSSK), UKM dipilih sebagai pakar linguistik Bahasa Melayu kajian ini. Mereka menerima latihan secara formal semasa diperingkat sarjana muda melalui subjek Analisis Wacana Bahasa Melayu dan diperkukuhkan secara langsung diperingkat Sarjana melalui subjek Seminar Sintaksis Bahasa Melayu dan subjek Seminar Semantik Bahasa Melayu. Tiga pakar linguistik bahasa Melayu ini menentukan ayat yang mengandungi isi utama berita dan seterusnya menghasilkan ringkasan rujukan secara ekstraktif yang mirip dengan ringkasan isi utama bagi setiap dokumen berita dalam korpus tersebut. Ketiga-tiga pakar ini menghasilkan satu ringkasan tajuk berita bagi setiap dokumen genre berita secara manual. Pakar perlu mematuhi tiga syarat iaitu ringkasan dilakukan secara pengekstrakan, teknik pemilihan perkataan secara *select-word-in-order* dan perubahan morfologi perkataan dibenarkan seperti yang ditetapkan dalam eksperimen (Zajic, Dorr & Schwartz, 2005). Seratus empat puluh ringkasan tajuk berita ini dianggap sebagai ringkasan yang mengandungi idea utama setiap dokumen genre berita dalam korpus ini.

Hasil dari cerapan pakar ini dianalisis dan dibandingkan dengan hasil yang diperoleh oleh Zajic, Dorr dan Schwartz (2005) dan dibincangkan secara terperinci dalam bahagian keputusan. Hasil tersebut merupakan ciri sistem bahasa Melayu yang perlu dipertimbangkan dalam teknik ringkasan tajuk berita.

KEPUTUSAN

Secara keseluruhannya, 96.4% ringkasan tajuk berita yang dihasilkan oleh kumpulan pakar mematuhi syarat eksperimen dan 3.6% atau 5 daripada 140 ringkasan *headline* pula tidak mematuhi syarat eksperimen yang telah ditetapkan. Hasil analisis struktur binaan dokumen berita bagi 140 dokumen berita dalam korpus dokumen berita kajian ini menunjukkan kesemua dokumen berita dibina berdasarkan sistem piramid terbalik (Md Salleh, 1985). Penggunaan sistem piramid terbalik menunjukkan susunan perenggan dalam binaan struktur dokumen berita dibuat berdasarkan tahap kepentingan tinggi ke rendah. Dua faktor yang menyebabkan sistem piramid terbalik digunakan adalah untuk (i) memudahkan pembaca memahami kandungan dokumen berita, dan (ii) memudahkan sidang pengarang untuk menyesuaikan dokumen berita berdasarkan saiz ruangan cetakan secara fizikal atau paparan dalam portal rasmi agensi penerbitan akhbar. Ini jelas menunjukkan struktur binaan dokumen berita mempengaruhi kedudukan isi utama penulisan dokumen berita Bahasa Melayu.

Bahagian permulaan dalam struktur binaan dokumen berita dilihat lebih berpotensi mengandungi isi utama penulisan dokumen berita.

Teknik pemilihan perkataan secara *select-word-inorder* adalah syarat yang tidak dipatuhi oleh kelima-lima ringkasan tajuk berita tersebut. Ini menyebabkan susunan perkataan dalam ringkasan tajuk berita tidak selari dengan susunan perkataan dalam ayat terpenting seperti yang terdapat di dalam Jadual 2. Susunan perkataan memainkan peranan penting bagi memastikan makna sebenar ayat terpenting dapat dipelihara.

JADUAL 2. Ringkasan Tajuk Berita Tidak Mematuhi Syarat Eksperimen

DOKUMEN	AYAT PENTING	RINGKASAN TAJUK BERITA
BM077	Transformasi ¹ kepada sistem ² pendidikan ³ negara ⁴ perlu ⁵ dilaksanakan	Sistem ² Pendidikan ³ Negara ⁴ Perlu ⁵ Transformasi ¹
BM099	... sukan SEA ke-26 yang memenangi ¹ pingat ² emas ³ selepas menewaskan Indonesia kelmarin, menerima ganjaran ⁴ wang tunai RM30,000 ⁵ , ...	Ganjaran ⁴ RM30,000 ⁵ Menang ¹ Pingat ² Emas ³
BM119	Bagi melahirkan ¹ lebih ramai ² usahawan ³ tempatan berdaya saing, Kolej ⁴ Komuniti ⁵ kini ...	Kolej ⁴ Komuniti ⁵ Lahir ¹ Ramai ² Usahawan ³
BM131	Kementerian ¹ Pelajaran ² menyambut ³ baik ⁴ hasrat Universiti Kebangsaan Malaysia (UKM) ⁵ dengan ⁶ kerjasama ⁷ Permata ⁸ Negara ⁹ membangunkan ...	Kementerian ¹ Pelajaran ² Sambut ³ Baik ⁴ Kerjasama ⁷ UKM ⁵ Dengan ⁶ Permata ⁸ Negara ⁹
BM135	India ¹ mengakui mereka semakin menghormati Malaysia ² dan lebih berwaspada ³ apabila berhadapan ⁴ tuan rumah ...	India ¹ Berwaspada ³ Berhadapan ⁴ Malaysia ²

Sebagai contoh untuk dokumen BM077, jika mengikut *select-word-inorder*, susunan perkataan haruslah mengikut urutan (1)'transformasi', (2)'sistem', (3)'pendidikan', (4)'negara', (5)'perlu'. Tetapi urutan berbeza digunakan dalam penghasilan ringkasan headline. Sehubungan dengan itu, lima ayat dalam Jadual 2 tidak mematuhi syarat *select-word-inorder*. Hasil analisis ke atas ringkasan tajuk berita yang dihasilkan oleh kumpulan pakar mendapati tiga ciri sistem bahasa Melayu dalam dokumen genre berita telah dikenal pasti. Tiga ciri tersebut adalah calon ayat terpenting, takrifan akronim dalam ayat dan saiz ringkasan.

Ringkasan rujukan isi utama yang dijana oleh teknik ringkasan isi utama berdasarkan ciri kata yang betul adalah seperti dalam Jadual 3 berikut. Jadual 3 menunjukkan dokumen ID BM071 mempunyai dua ayat terpenting dengan pemberat yang telah ditentukan. Dua ayat terpenting ini dijadikan asas bagi penjana ringkasan tajuk utama.

JADUAL 3. Ringkasan Tajuk Berita Yang Mematuhi Syarat Eksperimen

Kategori	Dokumen
Dokumen Asal BM071	KUALA LUMPUR: Empat pingat emas Olimpik London 2012 dalam jangkauan atlit negara. Kenyataan itu diluahkan oleh Menteri Belia dan Sukan, Datuk Seri Ahmad Shabery Cheek yang menegaskan sungguhpun Malaysia masih belum pernah meraih sebutir emas pada temasya berprestij itu, beberapa atlit dalam sukan badminton, berbasikal dan terjun mampu melakukannya. Ahmad Shabery merujuk kepada prestasi pemain badminton nombor satu dunia, Lee Chong Wei dan pasangan beregu utama, Koo Kien Keat-Tan Boon Heong serta jaguh berbasikal trek, Azizulhasni Awang dan penerjun, Pandelega Rinong yang disifatkan sebagai juara dalam acara masing-masing atau hanya sedikit di belakang juara sebenar. "Empat pingat emas ini bukanlah sasaran di Olimpik London tetapi apa yang saya maksudkan itu bersandarkan kepada keupayaan semasa beberapa atlit

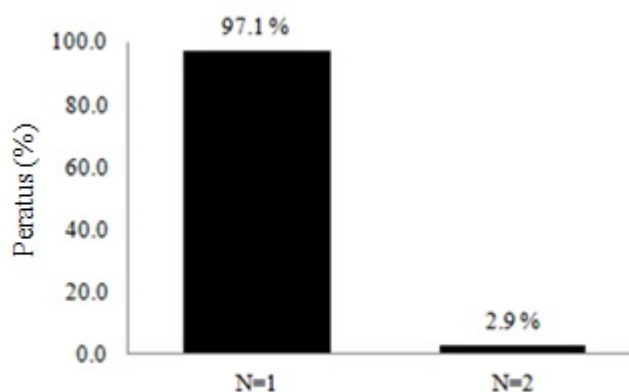
negara dalam sukan tertentu,” katanya pada majlis memeterai Perjanjian Persefahaman antara Majlis Sukan Negara (MSN) dengan 100PLUS dan Pelancaran Detik 500 Hari Program 'Road To London' di Kompleks Sukan Negara di Bukit Jalil, semalam

	ID BM071		
	No Ayat	Ayat	Pemberat
Dua ayat terpenting berserta pemberat	1	KUALA LUMPUR : Empat pingat emas Olimpik London 2012 dalam jangkauan atlit negara .	0.7000
	2	Kenyataan itu diluahkan oleh Menteri Belia dan Sukan , Datuk Seri Ahmad Shabery Cheek yang menegaskan sungguhpun Malaysia masih belum pernah meraih sebutir emas pada temasya berprestij itu , beberapa atlit dalam sukan badminton , berbasikal dan terjun mampu melakukannya .	0.1622

Ringkasan Tajuk Berita Empat Pingat Emas Olimpik London 2012 Jangkauan Atlit Negara

CALON AYAT TERPENTING

Penelitian ke atas 140 ringkasan tajuk berita yang telah dihasilkan oleh kumpulan pakar menunjukkan dua ayat pertama sahaja dipilih sebagai ayat terpenting yang mengandungi idea utama penulisan dokumen genre berita. Rajah 1 menunjukkan 97.1% ayat terpenting berada pada ayat pertama dan 2.9% ayat terpenting berada pada ayat kedua. Hasil eksperimen Zajic, Dorr dan Schwartz (2005), bagaimana pun menunjukkan idea utama penulisan dokumen genre berita bahasa Inggeris berada pada perenggan pertama atau 50 perkataan permulaan bagi setiap dokumen genre berita.



RAJAH 1. Bilangan ayat terpenting

POTENSI TAKRIFAN AKRONIM DALAM AYAT

Penelitian terhadap akronim dalam 140 dokumen genre berita dan ringkasan tajuk berita telah dilakukan dan didapati 45% dokumen genre berita mengandungi akronim pada ayat pertama, kedua atau kedua-duanya dan 37% ringkasan tajuk berita juga mengandungi akronim. 98% ringkasan tajuk berita tersebut dihasilkan daripada ayat terpenting yang mengandungi takrifan akronim. Ini menunjukkan ayat yang mengandungi takrifan akronim lebih berpotensi dianggap ayat terpenting dalam dokumen genre berita bahasa Melayu. Fitur ini juga tidak dibincangkan dalam eksperimen yang dilakukan oleh Zajic, Dorr dan Schwartz, (2005).

Analisis data menunjukkan jenis kata akronim dan kata mewakili nama individu yang wujud dalam dua ayat pertama lebih cenderung dipilih sebagai kata dalam ringkasan rujukan tunggal. Sejumlah 34% atau 48 ringkasan rujukan tunggal yang dihasilkan oleh kumpulan

pakar linguistik Bahasa Melayu didapati mengandungi kata berjenis akronim yang dipilih daripada dua ayat pertama dokumen berita. Sejumlah 15% atau 21 ringkasan rujukan tunggal pula didapati mengandungi kata mewakili nama individu yang dipilih daripada dua ayat pertama dokumen berita. Rajah 2 dan 3 masing-masing menunjukkan kecenderungan kata berjenis kata akronim dan kata mewakili nama individu dipilih daripada dua ayat pertama dokumen berita bagi membentuk ringkasan rujukan tunggal.

Dokumen : BM093
Model Ringkasan Isi Utama : Dasar MBMMBI Harus Dihormati
KUALA LUMPUR 29 Oktober – Keputusan kerajaan untuk memperkenalkan dasar Memartabat Bahasa Melayu dan Memperkukuh Bahasa Inggeris (MBMMBI) dan memansuhkan Pengajaran dan Pembelajaran Sains dan Matematik dalam Bahasa Inggeris (PPSMI) harus dihormati dan diterima baik semua pihak, kata Tan Sri Muhyiddin Yassin.¹

Rajah 2. Kecenderungan pemilihan kata berjenis kata akronim

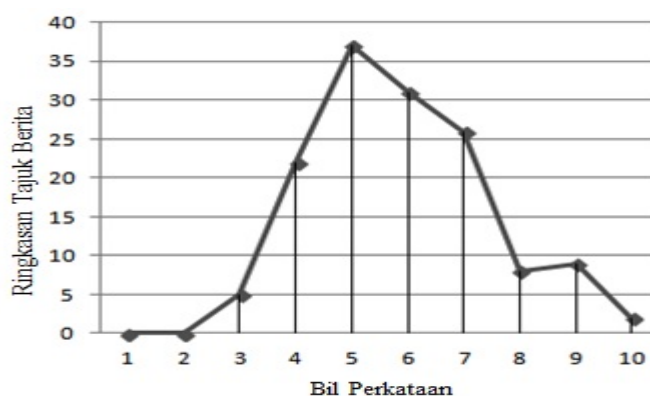
Dokumen : BM105
Model Ringkasan Isi Utama : Azizulhasni Awang Sandaran Utama Negara
KUALA LUMPUR 30 Mac - Pemenang pingat perak pada Kejohanan Dunia, Azizulhasni Awang akan menjadi sandaran utama negara untuk meraih tiga emas pada Kejohanan Berbasikal Asia (ACC) di Dubai bermula 10 April ini.¹

RAJAH 3. Kecenderungan pemilihan kata mewakili nama individu

Berdasarkan hasil analisis aras permukaan teks, tiga kesimpulan dapat dibuat iaitu (i) dokumen berita perlu memiliki sifat lengkap menyebabkan taburan frekuensi kata dalam dokumen berita menjadi hampir sekata, (ii) kata dipermulaan ayat mempunyai tahap kepentingan lebih tinggi daripada kata seterusnya dalam ayat yang sama, dan (iii) kata berjenis kata akronim dan kata mewakili nama individu yang wujud dalam dua ayat pertama lebih cenderung dipilih sebagai kata dalam ringkasan rujukan tunggal.

SAIZ RINGKASAN TAJUK BERITA

Analisis data menunjukkan purata saiz ringkasan tajuk berita yang telah dihasilkan oleh kumpulan pakar adalah 5.86 perkataan. Rajah 4 menunjukkan saiz bagi 140 ringkasan tajuk berita yang telah dihasilkan oleh kumpulan pakar. Oleh itu, saiz ringkasan tajuk berita ideal bagi dokumen genre berita bahasa Melayu adalah enam perkataan. Jadual 4 menunjukkan tiga contoh ringkasan tajuk berita ideal (enam perkataan). Saiz ini bersamaan dengan saiz piawai bagi penghasilan tajuk berita yang dipraktikkan dalam penerbitan akhbar arus perdana bahasa Melayu di Malaysia (Saiful Nujaimi, 2009).



RAJAH 4. Saiz ringkasan perkataan yang dihasilkan oleh kumpulan pakar

JADUAL 4. Contoh ringkasan tajuk berita bersaiz enam perkataan

Berita	Tajuk Berita Ideal
<p>ID : BM029 KUALA LUMPUR 7 Januari - Rizab antarabangsa Bank Negara Malaysia (BNM) berjumlah RM328.6 bilion (bersamaan dengan US\$106.5 bilion) pada 31 Disember 2010. Dalam satu kenyataan di sini hari ini, BNM berkata, kedudukan rizab mencukupi untuk membiayai 8.5 bulan import tertangguh dan 4.1 kali hutang luar negeri jangka pendek. "Paras rizab pada 31 Disember 2010 telah mengambil kira peningkatan nilai rizab disebabkan oleh pelarasan suku tahunan penilaian semula pertukaran asing berikutan pengukuhan kebanyakan mata wang utama berbanding dolar pada suku keempat," katanya. BNM berkata, untuk keseluruhan tahun 2010, rizab antarabangsa menurun sedikit sebanyak RM2.7 bilion kepada RM328.6 bilion (akhir-2009: RM331.3 bilion). Dalam tahun berkenaan, akaun semasa kekal mencatatkan lebihan dan terdapat aliran masuk pelaburan langsung asing dan dana portfolio, katanya.</p>	<p>Rizab Cukup Biaya Hutang Luar Negeri</p>
<p>ID : BM057 KUALA LUMPUR 8 Mei - Bank Negara Malaysia (BNM) telah mengambil kira kadar gaji minimum antara RM800 dan RM900 bagi unjuran inflasi antara 2.0 peratus dan 3.0 peratus pada tahun ini. Gabenor BNM, Tan Sri Zeti Akhtar Aziz berkata, kenaikan gaji minimum itu membabitkan sebahagian kecil daripada kos pengeluaran. "Berikutan itu, (kenaikan gaji) tidak akan menyumbang kepada kenaikan harga. Lagipun, harga masih sederhana. "Unjuran asal masih dikekalkan," katanya selepas merasmikan Persidangan Persatuan Institut-Institut Perbankan Asia Pasifik di sini hari ini. Malaysia akan melaksanakan gaji minimum sebanyak RM900 bagi Semenanjung dan RM800 di Sabah dan Sarawak, selepas enam bulan akta berkaitan diwartakan. Mengulas mengenai perkara itu, Zeti berkata, sebanyak 100 buah negara telah melaksanakan gaji minimum. Tambah beliau, kenaikan gaji itu adalah positif kerana ia akan menyumbang kepada perbelanjaan. "Kenaikan (gaji) juga akan menyumbang kepada pertumbuhan," ujarnya.</p>	<p>BNM Ambil Kira Kadar Gaji Minimum</p>
<p>ID : BM070 BANGI 10 Mac - Program Permata, yang memberi fokus kepada pembangunan dan pendidikan kanak-kanak, hari ini melangkah setapak lagi ke hadapan dengan penubuhan Yayasan Permata dengan modal permulaan sebanyak RM4.4 juta. Isteri Perdana Menteri, Datin Seri Rosmah Mansor yang juga Penaung Permata berkata, dengan penubuhan yayasan itu, semua program-program Permata yang dimulakan sejak 2007 iaitu Permata Negara, Permata Pintar, Permata Seni dan Permata Insan diharap dapat dilaksanakan dengan lebih terancang dan tersusun di bawah naungan satu yayasan, tanpa bergantung sepenuhnya kepada pembiayaan kerajaan. Beliau berkata, dana itu juga mungkin akan disalurkan dalam bentuk biasiswa pendidikan kepada pelajar Permata Pintar yang terpilih. "Saya percaya masih banyak lagi perkara yang boleh kita lakukan untuk kepentingan kanak-kanak di negara kita, dan dengan adanya dana yang mencukupi, bolehlah kita merancang program-program yang baik untuk kanak-kanak," katanya semasa majlis pelancaran Yayasan Permata di Kompleks Permata Pintar, Universiti Kebangsaan Malaysia (UKM) di sini hari ini.</p>	<p>Program Permata Fokus Pembangunan Pendidikan Kanak-kanak</p>

PERBINCANGAN

CALON AYAT TERPENTING

Hasil eksperimen ini adalah selari dengan amalan penulisan berita dalam akhbar arus perdana bahasa Melayu di Malaysia. Menurut Saiful Nujaimi (2009), idea utama penulisan berita bahasa Melayu berada pada bahagian pendulu berita dan tajuk berita akan menjadi sebagai

ringkasan kepada pendulu berita. Pendulu berita ditakrifkan sebagai ayat pertama pada perenggan pertama bagi sebuah dokumen berita bahasa Melayu (Md Salleh, 1985).

Hasil analisis ini menunjukkan sejumlah 97% atau 136 ringkasan rujukan tunggal yang dihasilkan oleh kumpulan pakar terdiri daripada kata yang dipilih daripada ayat pertama dan selebihnya dipilih daripada ayat kedua. Oleh itu, dapat disimpulkan dua ayat pertama dalam berita lebih berpotensi mengandungi isi utama penulisan berita dan kata yang wujud dalam dua ayat pertama juga lebih cenderung dipilih sebagai kata dalam ringkasan rujukan tunggal. Oleh itu, dua ayat pertama dalam dokumen genre berita bahasa Melayu adalah calon sesuai bagi ayat terpenting.

POTENSI TAKRIFAN AKRONIM DALAM AYAT

Akronim merupakan kependekan yang terbentuk daripada gabungan huruf pertama atau suku kata beberapa perkataan yang boleh diujarkan sebagai satu perkataan (Nik Safiah Karim et al., 2010). Dalam sistem bahasa Melayu, akronim dianggap kata tunggal yang diterbitkan hasil daripada penggabungan dan penyingkatan beberapa perkataan.

Kata berjenis kata akronim merupakan fitur kata ketiga yang dikenal pasti signifikan dengan isi utama penulisan dokumen berita Bahasa Melayu. Kata akronim lebih cenderung dipilih sebagai kata dalam ringkasan rujukan tunggal yang dihasilkan oleh kumpulan pakar linguistik Bahasa Melayu. Ini disokong oleh hasil analisis yang menunjukkan 34% ringkasan rujukan tunggal mengandungi kata berjenis kata akronim yang dipilih daripada dua ayat pertama. Ini menunjukkan kata berjenis akronim yang wujud dalam dua ayat pertama lebih signifikan dengan isi utama penulisan berita. Sebagai kesimpulannya, kata berjenis kata akronim adalah ciri kata yang signifikan dengan isi utama penulisan dokumen berita Bahasa Melayu.

Dua ayat pertama bagi dokumen genre berita adalah calon yang sesuai bagi teknik ringkasan tajuk berita. Ayat yang mengandungi takrifan akronim pula mempunyai potensi yang besar untuk dipilih sebagai ayat yang terpenting. Ayat yang terpenting dianggap input sesuai kepada teknik ringkasan tajuk berita kerana ia mengandungi idea utama penulisan dokumen genre berita. Pertimbangan dua ciri ini diharapkan dapat menghasilkan teknik ringkasan tajuk berita yang lebih pintar.

SAIZ RINGKASAN TAJUK BERITA

Document Understanding Conference (DUC) dan kini dikenal sebagai *Text Analysis Conferences* (TAC) merupakan komuniti antarabangsa yang aktif dalam penyelidikan berkaitan dengan ringkasan teks automatik. Komuniti tersebut telah menetapkan saiz ringkasan tajuk berita yang ideal bagi dokumen versi bahasa Inggeris adalah 10 perkataan (Over, Dang & Harman, 2007). Penetapan saiz tersebut dibuat berdasarkan hasil eksperimen yang dilakukan oleh Zajic, Dorr & Schwartz, (2005). Saiz ringkasan tajuk berita ideal bagi dokumen berita bahasa Melayu adalah enam perkataan berbanding sepuluh perkataan bagi dokumen berita bahasa Inggeris. Pendapat ini selari dengan saiz piawai bagi penghasilan tajuk berita yang diamalkan dalam penerbitan berita bahasa Melayu di Malaysia.

KESIMPULAN

Berdasarkan eksperimen ini, dapat disimpulkan bahawa terdapat tiga fitur sistem bahasa Melayu yang perlu dipertimbangkan dalam teknik ringkasan tajuk berita bagi dokumen genre berita bahasa Melayu. Tiga fitur tersebut adalah calon ayat terpenting, potensi ayat mengandungi takrifan akronim dan saiz ringkasan tajuk berita ideal.

Teknik ringkasan isi utama berdasarkan ciri kata yang dibangunkan adalah salah satu daripada aplikasi berasaskan bahasa tabii khusus kepada teks Bahasa Melayu. Prinsip teknologi kecerdasan buatan yang digunakan diharapkan menjadi sumber inspirasi kepada penyelidik seterusnya. Sumbangan kajian ini diharapkan dapat membantu para penyelidik pembangunan aplikasi berasaskan bahasa tabii bagi menghasikan aplikasi yang lebih cekap. Cadangan perluasan yang diutarakan diharap memberi motivasi kepada para para penyelidik bagi meneruskan penyelidikan dalam pembangunan teknik ringkasan teks Bahasa Melayu khususnya ringkasan isi utama.

Langkah seterusnya dalam kajian ini adalah untuk membangunkan sistem ringkasan tajuk berita bagi dokumen genre berita bahasa Melayu yang menggunakan fitur yang diperolehi. Sistem tersebut diharapkan lebih pintar dan berupaya menjana ringkasan tajuk berita yang lebih mirip seperti dilakukan oleh manusia. Selain daripada aspek resolusi anafora untuk Bahasa Melayu juga perlu dipertimbangkan bagi menghasilkan ringkasan teks yang lebih bermutu (Noorhuzaimi Karimah et al., 2012).

RUJUKAN

- Alicja P., Vania D. & Katja M. (2017). Automatic Extraction of News Values from Headline Text. Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics Valencia, Spain, 64–74.
- Alireza B. & Moses S. (2013). Headlines in Newspaper Editorials: A Contrastive Study. *SAGE Open, Vol. April-June(2013)*, 1-10.
- Alotaiby, F., Foda, S. & Alkharashi, I. (2011). A New Approach to HMM-Based Automatic Headline Generation. In Proceedings of the Computational Linguistics-Applications Conference Jachranka, 2011.
- Azmi, A. H. & Al-Thanyyan, S. (2012). A Text Summarizer for Arabic. *Computer Speech and Language, Vol. 26(2012)*, 260-273.
- Brooks, B. S. (1996). *News Reporting and Writing (5th ed.)* New York: St Martin's Press.
- Dorr, B., Zajic, D. & Schwartz, R. (2003). Cross-Language Headline Generation for Hindi. *ACM Transactions on Asian Language Information Processing, Vol. 2(3)*, 270-289.
- Dauzidia, F. S. & Lapalme, G. (2004). Lakhas, An Arabic Summarization System. Proceedings of the Document Understanding Conference (DUC 2004).
- Gunawan, D, Pasaribu, A, Rahmat, RF & Budiarto, R. (2017). Automatic Text Summarization for Indonesian Language Using TextTeaser. *IOP Conference Series: Materials Science and Engineering, Vol. 190(1)*.
- Hishamudin Isam & Norsimah Mat Awal. (2011). Analisis Berasaskan Korpus Dalam Menstruktur Semula Kedudukan Makna Teras Leksikal Setia. *GEMA Online® Journal of Language Studies, Vol. 11(1)*, 143-158.
- Julian, H. & Stanley, J. (1965). *The Complete Reporter*. New York: MacMillan Publishing Co. Inc.
- Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development, Vol. 2(2)*, 159-165.
- Md Salleh Kassim (1985). *Kewartawan: Teori & Praktis*. Dewan Bahasa dan Pustaka.
- Nik Safiah Karim, Farid M. Onn, Hashim Haji Musa & Abdul Hamid Mahmood. (2010). *Tatabahasa Dewan Edisi Ketiga*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Saiful Nujaimi Abdul Rahman (2009). *Kewartawan Malaysia: Praktis & Cabaran Dalam Era Revolusi Digital*. Kuala Lumpur: Prentice Hall.
- Tengku Mohd Tengku Sembok (2007). *Bahasa, Kecerdasan dan Makna Sekitar Capaian Maklumat*. Bangi: Penerbit Universiti Kebangsaan Malaysia.

- Shahrul Azman Mohd Noah, Nazlia Omar & Amru Yusrin Amruddin (2015). Evaluation of Lexical-Based Approaches to the Semantic Similarity of Malay Sentences. *Journal of Quantitative Linguistics*. Vol. 22(2), 135-156.
- Kaikhah, K. (2004). Automatic Text Summarization with Neural Networks. Second IEEE International Conference on Intelligent System, Hamburg-Harburg, Germany: 40-44.
- Kyoomarsi, F., Khosrawi, H., Eslami, E. & Dehkordy, P. K. (2008). Optimizing Text Summarization Based on Fuzzy Logic. Proceedings of Seventh IEEE/ACIS International Conference on Computer Science and Information Science, University of Shahid Bahonar Kerman, United Kingdom.
- Norshuhani Zamin & Arina Ghani. (2011). Summarizing Malay Text Documents. *World Applied Sciences Journal*. Vol. 12, 39-46.
- Mazdak, N. (2004). FarsiSum – A Persian Text Summarize. Unpublished M.S. Thesis, Department of Linguistics, Stockholm University.
- Noorhuzaimi Karimah Mohd Noor, Shahrul Azman Mohd Noah, Mohd Juzaidin Ab Aziz & Mohd Pouzi Hamzah (2012). Malay anaphor and antecedent candidate identification: a proposed solution. In Pan JS., Chen SM., Nguyen N.T. (Eds.), *Intelligent Information and Database Systems. ACIIDS 2012 Lecture Notes in Computer Science, vol 7198* (pp. 141-151). Springer, Berlin, Heidelberg.
- Nor Hashimah Jalaluddin & Ahmad Harith Syah. (2009). Penelitian Makna Imbuhan – Pen Dalam Bahasa Melayu: Satu Kajian Rangka Rujuk Silang. Satu Kajian Rangka Rujuk Silang. *GEMA Online® Journal of Language Studies*. Vol.9(2), 57-72.
- Over, P, Dang, H. & Harman, D. (2007). DUC in Context. *Information Processing and Management*. Vol. 43(2007), 1506-1520.
- Shen, D. (2009). *Text Summarization*. In. Liu, L. & Ozs, M. T. Encyclopedia of Database Systems. New York: Springer Science-Business Media.
- Shamsfard, M., Akhavan, T. & Erfani Jourabchi, M. (2009). PARSUMIST: A Persian Text Summarization. Proceedings of the 5th IEEE International Conference On Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'09).
- Steinberger, J. & Jezek, K. (2009). Text Summariazation: An Old Challenge and New Approaches. In. Ajit Abraham, Aboul-Ella Hassanien, Andre Ponce de Leon, F. De Carvalho & Vaclav Snasel. *Foundations of Computational Intelligence Volume 6, Data Mining*. Berlin: Springer-Verlag.
- Suraya Alias, Siti Khaotijah Mohammad & Hoon, G. K. (2018). A Text Representation Model Using Sequential Pattern-growth Method. *Pattern Anal Applic*. Vol. 21(1), 233-247.
- Zhou, L. & Hovy, E. (2003). Headline summarization at ISI. Proceedings of the Document Understanding Conference 2003 (DUC 2003).
- Zhou, L. & Hovy. E. (2004). Template-filtered headline summarization. Proceedings of the Association for Computational Linguistics. Workshop on Text Summarization Branches Out, 56-60.
- Zajic, D., Dorr, B. & Schwartz, R. (2002). Automatic headline generation for newspapers stories. Proceedings of the Document Understanding Conference 2002 (DUC 2002).
- Zajic, D., Dorr, B. & Schwartz, R. (2005). Headline Generation for Written and Broadcast News. Technical Report UMIACS-TR-2005-07, University of Maryland, USA.

PENULIS

Shahrul Azman Mohd Noah mendapat PhD daripada Shieffield Universiti United Kingdom. Merupakan Profesor di Fakulti Teknologi dan Sains Maklumat UKM. Bidang kepakaran beliau adalah capaian maklumat dan kepintaran buatan.

Nazlena Mohamad Ali merupakan felo penyelidik kanan dan Profesor Madya di Institut Informatik Visual, UKM. Mendapat PhD daripada Dublin City University, Ireland pada tahun 2009. Bidang kepakaran beliau adalah Antaramuka Manusia-Mesin.

Mohd Sabri Hasan mendapat PhD daripada Fakulti Teknologi dan Sains Maklumat, UKM pada tahun 2015. Bidang penyelidikan beliau adalah komputeran linguistik.