Penjanaan Ringkasan Isi Utama Berita Bahasa Melayu berdasarkan Ciri Kata

Shahrul Azman Mohd Noah
shahrul@ukm.edu.my
Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia

Nazlena Mohamad Ali
nazlena.ali@ukm.edu.my
Institut Informatik Visual,
Universiti Kebangsaan Malaysia

Mohd Sabri Hasan
sabri@kabinet.gov.my
Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia

ABSTRAK

Teknik ringkasan isi utama merupakan satu proses penyulingan maklumat penting daripada wacana untuk menghasilkan satu ayat tunggal yang mewakili isi utama penulisan. Dalam konteks wacana Bahasa Melayu, kajian bidang ini terlalu sedikit dan tertumpu kepada kaedah penterjemahan mesin. Kajian ini dibahagikan kepada tiga fasa iaitu analisis korpus wacana berita, pembangunan teknik ringkasan isi utama dan penilaian kualiti hasil ringkasan. Kajian bertujuan untuk membangunkan teknik ringkasan isi utama dengan menggabungkan kaedah statistik dan linguistik. Kaedah statistik digunakan untuk menentukan kata signifikan dan ayat terpenting berdasarkan konsep pemberat. Kaedah linguistik pula digunakan untuk meningkatkan ketepatannya. Korpus wacana berita Bahasa Melayu terdiri daripada 140 wacana berita berserta ringkasan rujukan tunggal. Hasil analisis korpus wacana berita mendapati isi utama penulisan berita dapat ditentukan berdasarkan empat ciri iaitu lokasi kedudukan kata dalam ayat, kedudukan dua ayat pertama wacana berita, kata berjenis akronim dan kata mewakili nama individu. Kata signifikan dengan isi utama penulisan teks ditentukan berdasarkan nilai pemberat kata. Nilai ditentukan dengan menggabungkan nilai frekuensi kata dalam dokumen dan kedudukan kata dalam ayat. Dua ayat pertama dalam dokumen berita Bahasa Melayu dikenalpasti sebagai calon ayat terbaik bagi pengecaman ayat terpenting. Hasil penilaian menunjukkan peratus min ketepatan pengecaman ayat terpenting adalah 82.9% dan min kualiti ringkasan isi utama yang dijanakan masing-masing ialah kejituan (0.3194), dapatan semula (0.5656), skor-F (0.4012), ROUGE-N (0.5656), ROUGE-L (0.3392), ROUGE-W (0.1186) dan ROUGE-S (0.1232). Kesimpulannya pertimbangan faktor bahasa dalam pembangunan teknik ringkasan isi utama mampu menghasilkan ringkasan yang berkualiti daripada aspek bahasa dan darjah ketepatan yang lebih baik.

Kata Kunci: peringkasan teks; kaedah tanpa seliaan; ciri kata; berita Bahasa Melayu

eISSN: 2550-2131
ISSN: 1675-8021

Generation of News Headline for Malay Language based on Term Features

ABSTRACT

Headline generation is an information extraction process to generate a single sentence that represents the content of a text. In Malay language context, research in this area is limited to machine translation approaches. This study is divided into three phases: analysis of news discourse, development of headline generation technique and evaluation of the quality of generated headlines. The study aims to develop headline using statistical and linguistic methods. The statistic method used to identify significant words and sentences based in term weighting approach. The linguistic method is used to increase its preciseness. 140 news and their corresponding headlines model were constructed. Analysis of the news collection shows that the main idea of written text can be identified based on four characteristics: word location in sentences, sentence location in texts, acronym word types and words that represent the person name. Significant words with main idea of written text are determined based on the words weighted values. The values are determined by combining the frequency of words and word location in sentences. The content of the first two sentences are suitable candidates for recognising important sentences in text. Results showed that mean percentage for important sentence recognition 82.9%, mean quality of generated headlines are 0.3194 (precision), 0.5656 (recall), 0.4012 (F-measure), 0.5656 (ROUGE–N), 0.3392 (ROUGE–L), 0.1186 (ROUGE–W) and 0.1232 (ROUGE–S). In conclusion, the consideration of language factors in headline generation technique is capable of producing quality headlines with higher degree of fidelity as compared to the compared benchmarks.

Keywords: text summarisation; unsupervised approach; term features; Malay news article

PENGENALAN

Setiap bahasa mempunyai tatabahasa yang membolehkan ayat dicipta tanpa had bilangannya. Tatabahasa merupakan petua berdasarkan sesuatu teori tertentu untuk menghuraikan ayat dengan kaedah yang paling mudah, tepat dan lengkap (Karim et al. 2010). Tatabahasa menjadikan setiap bahasa tabii unik dan berbeza dalam mengungkapkan satu ayat yang sempurna. Sejumlah ayat tertentu yang dicipta dengan mematuhi tatabahasa membawa hanya satu tumpuan makna atau isi penulisan membentuk satu fokus pemikiran yang dikenali sebagai wacana. Wacana sama ada berbentuk tulisan atau lisan dibentuk daripada unit bahasa yang berkaitan, berturutan dan lengkap melahirkan satu kesatuan yang utuh untuk menyampaikan maklumat tentang sesuatu topik. Contoh wacana seperti berita, artikel saintifik, lirik lagu dan sebagainya.

Teknologi maklumat dan komunikasi adalah teknologi yang dianggap paling berjaya berbanding teknologi lain dan berupaya menguruskan data dengan sistematik. Melalui teknologi ini, data dikumpulkan melahirkan maklumat, maklumat dengan kepastian atau kebenaran pula melahirkan pengetahuan, pengetahuan menjadi asas tindakan manusia dalam penyelesaian masalah dan pemilihan tindakan yang terbaik berdasarkan pengetahuan terkumpul melahirkan manusia yang bijaksana (Sembok, 2007). Pada peringkat perkembangan teknologi maklumat, dominasi jenis data adalah berbentuk nombor. Namun, apabila gajet komunikasi mudah alih seperti telefon pintar menjadi popular, teks menjadi jenis data yang paling dominan dalam persekitaran digital. Aplikasi media sosial adalah penyumbang utama data berbentuk teks sama ada berstruktur atau tidak berstruktur. Kajian bidang linguistik telah meningkat sejajar dengan perkembangan teknologi. Kajian melibatkan

teknologi dan bahasa Melayu dijalankan oleh beberapa penyelidik dalam aspek penyelenggaraan data terancang, sistematik dan berkesan (Hishamudin & Norsimah, 2011; Nor Hashimah & Ahmad, 2009; Alshalabi, Sabrina & Nazlia, 2017). Kajian lain pula berkaitan penentuan fitur bagi pengekstrakan tajuk berita, khusus kepada dokumen genre berita akhbar bahasa Melayu (Shahrul Azman, Nazlena & Mohd Sabri, 2018).

Teknik ringkasan tajuk berita berpotensi mengurangkan masalah kebanjiran maklumat yang berlaku dalam kepesatan teknologi masa kini. Teknik ini berupaya menguruskan dokumen dalam kuantiti yang besar mengikut keperluan maklumat pengguna dengan mudah, cepat serta berupaya mengurangkan bebanan kognitif dalam penelitian dokumen relevan (Alireza & Moses, 2013; Gunawan, Pasaribu, Rahmat & Budiarto, 2017)

Kertas ini dimulakan dengan pengenalan dan kajian lepas. Bahagian seterusnya menjelaskan metodologi kajian yang digunakan dalam kajian ini. Bahagian ketiga melaporkan hasil analisis bagi mengenal pasti ciri-ciri signifikan dengan kedudukan isi utama penulisan wacana berita Bahasa Melayu. Teknik yang dibangunkan dijelaskan secara terperinci dalam bahagian keempat dan penilaian prestasi teknik tersebut dilaporkan dalam bahagian kelima. Bahagian keenam pula membincangkan hasil kajian ini dan makalah ini diakhiri dengan kesimpulan keseluruhan kajian ini.

KAJIAN LEPAS

TEKNIK RINGKASAN TEKS

Teknik ringkasan teks merupakan suatu proses penyulingan maklumat penting daripada dokumen untuk menghasilkan ringkasan bagi memenuhi keperluan pengguna dan tugas tertentu (Mani, 1999). Proses ini melibatkan kecekapan dan keberkesanan manusia dalam menentukan maklumat penting daripada dokumen tunggal atau koleksi dokumen secara ekstraktif atau abstraktif (Shen, 2009). Menurut Nenkova dan McKeown (2011), ringkasan yang dihasilkan atau output teknik ringkasan teks terbahagi kepada dua iaitu ringkasan berpetunjuk, dan ringkasan berinformasi. Ringkasan berpetunjuk merujuk kepada hasil ringkasan yang membolehkan pengguna mengenal pasti petunjuk maklumat penting yang wujud dalam dokumen. Ringkasan berinformasi pula merujuk kepada hasil ringkasan yang membolehkan pengguna memahami kandungan dokumen dengan hanya membaca hasil ringkasan sahaja. Selain itu, hasil ringkasan juga perlu memiliki tahap kesepaduan dan fideliti yang baik. Hasil ringkasan dengan tahap kesepaduan yang baik menggambarkan susunan kata dalam struktur binaan hasil ringkasan mematuhi peraturan tatabahasa dan sintaksis sesebuah sistem bahasa dan memastikan hasil ringkasan dapat dibaca dengan lancar dan difahami dengan mudah (Karim et al. 2010). Tahap fediliti yang baik pula menunjukkan hasil ringkasan berjaya memelihara makna seperti dalam dokumen asal (Gupta & Lehal, 2010).

Teknik ringkasan teks secara tradisinya merupakan sub-bidang bagi pemprosesan bahasa tabii (NLP). Menurut Kaikhah (2004), dua faktor keupayaan teknik ringkasan teks iaitu pertama teknik ini berupaya mengurus hasil carian yang relevan dengan keperluan maklumat pengguna dengan mudah, dan kedua teknik ini mampu mengurangkan kesukaran ketika pemilihan hasil carian yang relevan dalam kuantiti yang besar. Ketersediaan pelbagai dokumen digital dalam kuantiti yang besar dalam sistem capaian maklumat mendorong kepada penyelidikan aktif teknik ringkasan teks seperti kajian Alguliev (2011), Atkinson dan Munoz (2013), dan El-Fishawy et al. (2014).

Evolusi perkembangan teknik ringkasan teks bermula sejak 1950-an yang dipelopori oleh Luhn (1958) dan Edmudson (1969). Perkembangan teknik ini dipengaruhi oleh dua faktor iaitu pertama perkembangan kaedah pengekstrakan maklumat dan kedua penerokaan bidang kecerdasan buatan berasaskan pengetahuan. Perkembangan kaedah pengekstrakan

maklumat terutama dalam bidang capaian maklumat, sains perpustakaan dan automasi pejabat mempengaruhi perkembangan teknik ringkasan teks melalui penciptaan teknik atau kaedah menganalisis teks dalam kandungan dokumen pada aras permukaan teks. Penerokaan bidang kecerdasan buatan berasaskan pengetahuan melalui penyulingan maklumat penting secara mendalam juga mempengaruhi perkembangan teknik ini. Fasa baru kebangkitan teknik ringkasan teks bermula pada tahun 1990-an yang digerakkan oleh dua faktor iaitu (i) ketersediaan teks dalam format digital yang semakin bertambah, dan (ii) ketersediaan teks digital yang terlalu banyak dalam persekitaran Internet (Lin, 2009; Shen, 2009). Dua faktor ini adalah penyumbang kepada masalah kebanjiran maklumat yang berlaku dalam aplikasi berasaskan capaian maklumat seperti enjin gelintaran. Masalah ini menyebabkan pengguna aplikasi dihidangkan banyak pilihan berbanding dengan apa yang patut dibaca apabila berinteraksi dengan aplikasi tersebut (Gupta & Lehal 2010). Selain itu, pengguna juga menghadapi masalah pengulangan maklumat dalam output carian sehingga menyukarkan pemilihan output relevan berbanding dengan keperluan maklumatnya (Foong et al. 2010).

Menurut Lin dan Liang (2008), teknik ringkasan teks berupaya mengurangkan beban kognitif pengguna aplikasi berasaskan capaian maklumat ketika membaca dan meneliti hasil carian. Pengurang beban kognitif ini dilakukan melalui tiga kaedah iaitu meminimumkan pengulangan maklumat dalam hasil carian, memudahkan hasil carian yang relevan dalam kuantiti yang besar, dan membolehkan hasil carian yang relevan dikategorikan mengikut topik. Kini kebangkitan teknik ringkasan teks didorong oleh tren penggunaan peranti mudah alih seperti telefon pintar dan PDA yang meluas dalam masyarakat hari ini. Skrin paparan peranti mudah alih yang kecil menyebabkan pengguna menghadapi masalah ketika memaparkan maklumat berteks. Oleh itu, teknik ringkasan teks digunakan bagi menghasilkan ringkasan bagi mewakili maklumat berteks dalam peranti mudah alih (Lin, 2009).

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

TEKNIK RINGKASAN ISI UTAMA

Teknik ringkasan isi utama merupakan satu proses penyulingan maklumat penting daripada wacana atau dokumen secara ekstraktif atau abstraktif untuk menghasilkan satu ayat tunggal yang mewakili isi utama penulisan wacana atau dokumen (Hasan, 2015). Ringkasan yang terhasil daripada teknik ini mampu mengurangkan masa pengguna untuk membaca atau meneliti wacana asal. Ini kerana ringkasan tersebut hanya mengandungi maklumat penting dan pengulangan maklumat diminimumkan. Aplikasi teknik ini dalam capaian maklumat mampu mengurangkan bebanan kognitif pengguna kerana ia mampu menguruskan hasil capaian yang relevan dengan mudah dan mengurangkan kesukaran pemilihan hasil capaian relevan dalam kuantiti yang besar (Kaikhah, 2004).

Kajian pembangunan teknik ringkasan isi utama telah dimulakan oleh Banko et al. (2000) pada wacana berita Bahasa Inggeris menggunakan kaedah pemilihan model dengan kaedah seliaan data. Sejak itu, pelbagai teknik ringkasan isi utama dengan atau tanpa kaedah seliaan data dibangunkan untuk wacana berita Bahasa Inggeris (Zajic et al., 2002; Dorr et al., 2003; Zhou & Hovy, 2003; Zhou & Hovy, 2004; Soricut & Marcu, 2007; Xu et al., 2010). Perluasan teknik ini kepada wacana berita bukan Bahasa Inggeris telah dilakukan dengan

menggabungkan teknik sedia ada dengan kaedah penterjemahan mesin seperti kajian Dorr dan Zajic (2003) untuk wacana berita Bahasa Arab dan Dauzidia dan Lapalme (2004) untuk wacana berita Bahasa Hindi.

Gabungan teknik ringkasan isi utama sedia ada dan kaedah penterjemahan mesin didapati mempunyai tiga kelemahan utama iaitu kesukaran dalam menentukan ayat penting bagi wacana yang diterjemahkan (Dorr & Zajicn 2003), ketepatan penentuan ayat penting bergantung kepada ketepatan penterjemah mesin (Dauzidia & Lapalme, 2004) dan masalah fediliti yang menyebabkan makna sebenar tidak dapat dipelihara (Shamsfard et al., 2009). Kelemahan ini mendorong kepada usaha untuk membangunkan teknik ringkasan isi utama dengan mempertimbangkan jenis bahasa tabii yang membina wacana berita. Hasilnya, teknik ringkasan isi utama untuk wacana berita bukan bahasa Inggeris dihasilkan seperti kajian Alotaiby (2011) untuk wacana berita bahasa Arab, Lee dan Kim (2005) untuk wacana berita bahasa Korea, Muurisep dan Mutso (2005) untuk wacana berita bahasa Estonia, Daniel (2008) untuk wacana berita bahasa Belanda dan Vijayapal et al. (2011) untuk wacana berita bahasa Telugu.

Pembangunan teknik ringkasan teks untuk wacana berita Bahasa Melayu dimulakan oleh Zamin dan Ghani (2011) yang menggunakan teknik gabungan daripada SUMMARIST (Hovy & Lin, 1997) dan EstSum (Muurisep & Mutso, 2005). Output daripada teknik ini merupakan senarai ayat penting yang terdapat dalam wacana berita. Seperti dinyatakan sebelum, pendekatan penterjemahan mempunyai kelemahan khususnya dari aspek fideliti iaitu makna sebenar kandungan teks tidak terpelihara. Sehubungan dengan itu, kajian lanjutan berkaitan pembangunan teknik ringkasan isi utama untuk wacana berita Bahasa Melayu masih terbuka luas.

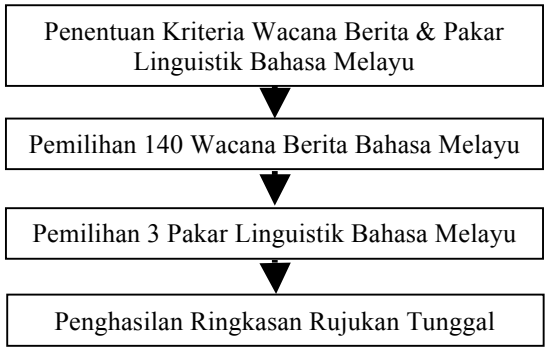
Kajian lepas menunjukkan pembangunan korpus dokumen berita Bahasa Inggeris bagi kajian teknik ringkasan isi utama hanya yang dibangunkan oleh Zajic et al. (2005) bagi dokumen berita Tipster telah dijelaskan secara terperinci. Tiada laporan pembangunan korpus dokumen berita Bahasa Melayu dilaporkan. Oleh itu, spesifikasi pembangunan korpus dokumen berita Tipster dijadikan penanda aras bagi pembangunan korpus dokumen berita Bahasa Melayu kajian ini. Korpus dokumen berita Tipster oleh Zajic et al. (2005) terdiri daripada 73 dokumen berita Bahasa Inggeris yang dipilih daripada koleksi dokumen berita Tipster. Dokumen berita tersebut merupakan berita daripada tiga genre berita iaitu harga komoditi, kronologi peristiwa dan aktiviti hiburan. Dengan menggunakan dua orang pakar, satu ringkasan rujukan tunggal bagi setiap dokumen berita dalam korpus tersebut dihasilkan secara manual. Ringkasan rujukan tunggal yang dihasilkan oleh dua pakar dinilai daripada empat aspek iaitu (i) pematuhan teknik pilih kata secara tertib kedudukannya dalam dokumen asal, (ii) bilangan kata dalam ringkasan rujukan tunggal, (iii) pemilihan nombor ayat terpenting; dan (iv) jarak maksimum kata yang dipilih sebagai kata dalam ringkasan rujukan tunggal daripada permulaan perenggan dokumen berita.

METODOLOGI KAJIAN

Kajian ini merupakan kajian kuantitatif dalam bidang ringkasan teks yang melibatkan pembangunan teknik ringkasan isi utama dengan mempertimbangkan faktor bahasa yang membina wacana berita. Kajian ini dilaksanakan bertujuan untuk membangunkan dan menilai teknik ringkasan isi utama bagi pengekstrakan isi utama penulisan wacana berita Bahasa Melayu.

Korpus wacana berita yang diguna dalam kajian ini dibangunkan berdasarkan kaedah pembangunan korpus wacana berita Inggeris untuk kajian teknik ringkasan isi utama yang dilaporkan dalam kajian Zajic et al. (2005). Walau bagaimanapun, beberapa penyesuaian telah dilakukan untuk diselaraskan dengan kajian ini. Hasil pembangunan korpus dokumen

berita ini merupakan 140 dokumen berita berserta ringkasan rujukan tunggal. Kaedah pembangunan korpus wacana berita ini diringkaskan seperti dalam Rajah 1.



Sumber: Hasan (2015)

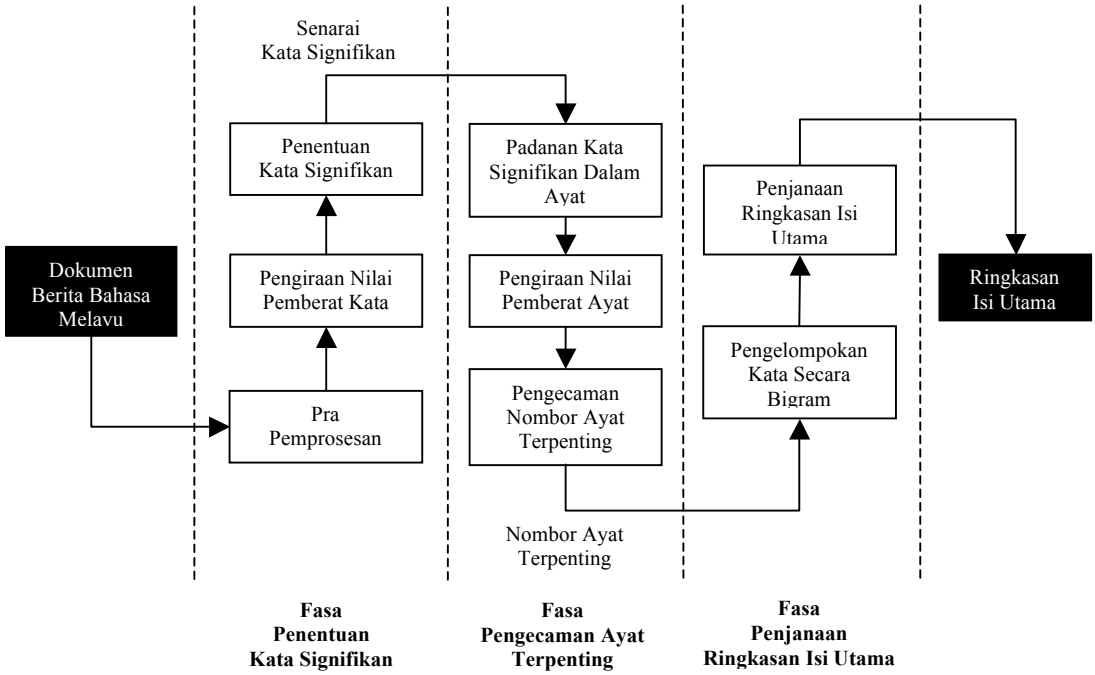
RAJAH 1. Kaedah pembangunan korpus wacana berita Bahasa Melayu

Kajian ini dibahagikan kepada tiga fasa iaitu analisis korpus wacana berita, pembangunan teknik ringkasan isi utama dan penilaian kualiti hasil ringkasan. Analisis korpus wacana berita bertujuan untuk mengenal pasti ciri-ciri signifikan daripada aspek bahasa yang menentukan kedudukan isi utama penulisan wacana berita Bahasa Melayu. Ciri-ciri yang dikenal pasti dikaji untuk dipertimbangkan dalam pembangunan teknik ringkasan isi utama. Seterusnya, teknik tersebut dinilai daripada aspek kualiti hasil ringkasan dengan membandingkan antara ringkasan yang dijanakan oleh teknik dengan ringkasan rujukan tunggal yang dihasilkan oleh kumpulan pakar.

TEKNIK RINGKASAN ISI UTAMA BERDASARKAN CIRI KATA

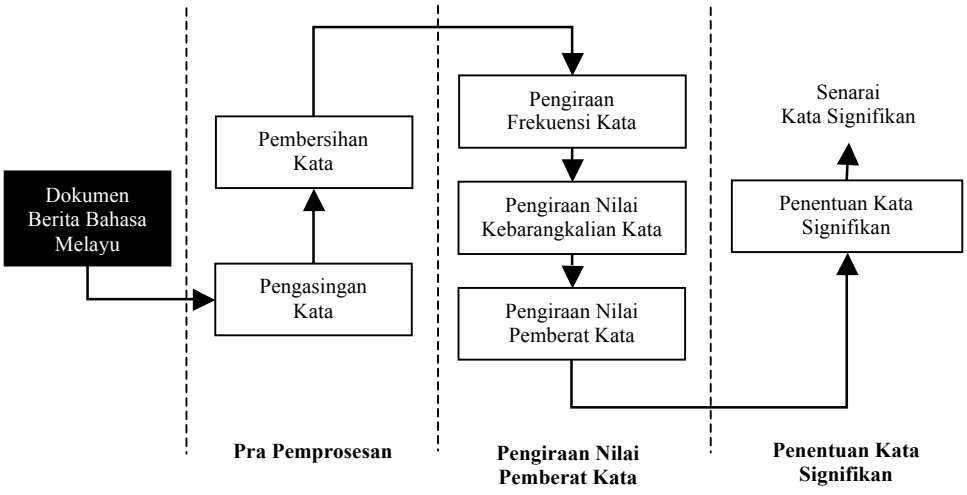
Kerangka kerja teknik ringkasan isi utama ini dibahagikan kepada tiga fasa iaitu fasa penentuan kata signifikan, fasa pengecaman ayat terpenting dan fasa penjanaan ringkasan isi utama seperti Rajah 2.

eISSN: 2550-2131
ISSN: 1675-8021



RAJAH 2. Kerangka kerja teknik ringkasan isi utama berdasarkan ciri kata

Fasa penentuan kata signifikan bertujuan untuk menentukan 10 kata signifikan dengan isi utama penulisan wacana berita tanpa melibatkan kaedah seliaan data. Penentuan kata signifikan ini melibatkan empat pertimbangan yang dikenal pasti semasa fasa analisis korpus wacana berita Bahasa Melayu. Fasa ini dibahagikan kepada tiga proses iaitu pra pemprosesan, pengiraan nilai pemberat kata dan penentuan kata signifikan seperti Rajah 3.



RAJAH 3. Fasa penentuan kata signifikan

Penentuan kata signifikan dibuat berdasarkan 10 kata yang mempunyai nilai pemberat tertinggi. Pengiraan nilai pemberat ditentukan berdasarkan idea Luhn (1958). Idea ini kerap digunakan dalam penentuan kata signifikan tanpa melibatkan kaedah seliaan data. Namun, dalam konteks wacana berita Bahasa Melayu dengan bilangan ayat kurang daripada 15 ayat atau 250 patah kata, idea tersebut kurang berkesan. Ini kerana frekuensi kata dalam dokumen menjadi hampir sekata. Oleh itu, kaedah frekuensi kata digabungkan dengan ciri kata signifikan yang dikenal pasti iaitu berdasarkan lokasi kedudukan kata dalam ayat. Persamaan bagi menentukan nilai pemberat bagi setiap kata adalah seperti persamaan 1, 2 dan 3.

$$w(k_i , ayat_j) = tf(k_i) + loc(k_i , ayat_j) \quad (1)$$

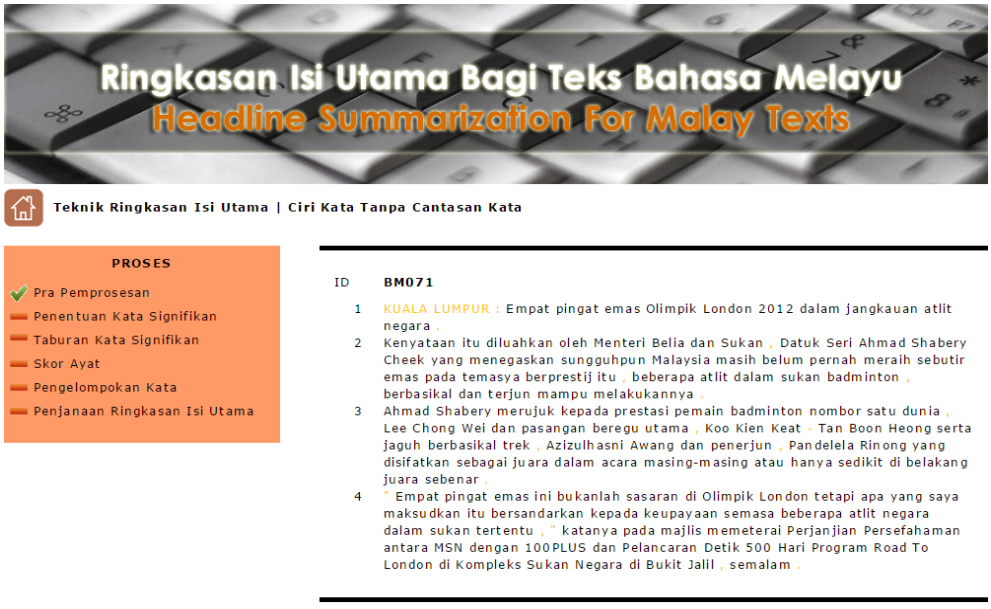
$$loc(k_i , ayat_j) = f(k_i) / m \quad (2)$$

$$f(k_i) = m - n + 1 \quad (3)$$

iaitu,

$w(k_i , ayat_j)$ = pemberat kata k_i dalam ayat j
$tf(k_i)$ = frekuensi kata k_i dalam ayat j
$loc(k_i , ayat_j)$ = lokasi kata k_i dalam ayat j
$f(k_i)$ = frekuensi kata k_i dalam ayat j
n = lokasi semasa kata k_i dalam ayat j
m = jumlah kata k_i dalam ayat j

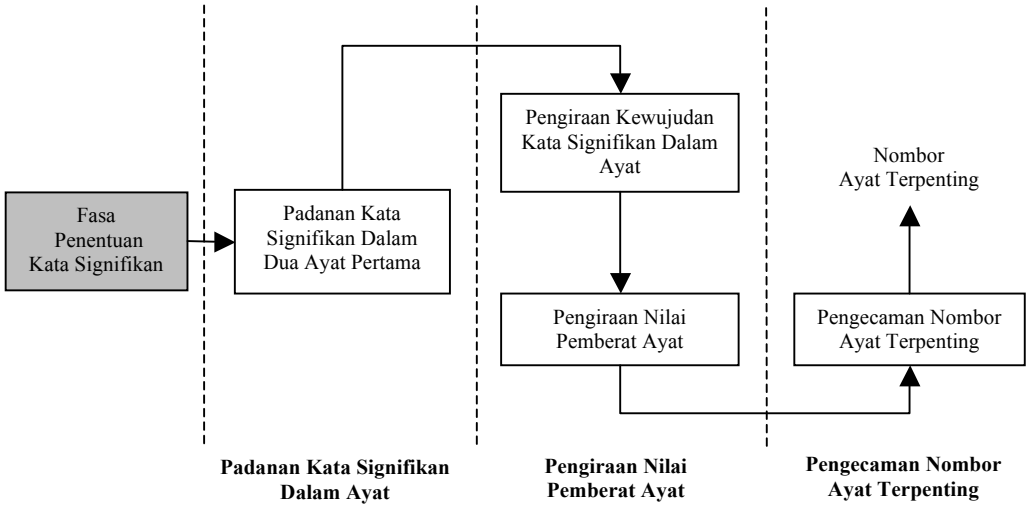
Pra pemprosesan merupakan proses yang bertujuan untuk menyediakan kata bagi proses pengiraan nilai pemberat kata dan penentuan kata signifikan. Proses ini dibahagikan kepada dua peringkat iaitu (i) pengasingan kata, dan (ii) pembersihan kata. Pengasingan kata adalah proses untuk mengasingkan ayat dalam dokumen berita kepada unit n-gram atau kata berdasarkan sempadan kata. Sempadan kata ditentukan berdasarkan ruang kosong yang wujud di antara dua kata dalam ayat. Proses pengasingan kata dilakukan ke atas 140 dokumen berita dan menghasilkan 22,774 unit n-gram. Seterusnya, unit n-gram yang terhasil melalui proses pembersihan kata. Proses pembersihan kata bertujuan untuk menghapuskan n–gram yang berjenis tanda baca dan kata henti. Proses pembersihan n-gram yang berjenis tanda baca melibatkan tanda baca noktah (.), tanda koma (,), tanda tanya (?), tanda seruan (!), dan tanda pengikat kata ("..."). Contoh pembersihan n-gram yang berjenis tanda baca bagi dokumen berita berlabel BM071 seperti Rajah 4.



RAJAH 4. Pra pemprosesan

RAJAH 5. Frekuensi bagi kata signifikan

Fasa pengecaman ayat terpenting bertujuan untuk mengecam nombor ayat terpenting berdasarkan 10 kata signifikan yang telah dikenal pasti dalam fasa sebelumnya (ditunjukkan dalam Rajah 5). Definisi ayat terpenting dalam kajian ini merujuk kepada ayat yang mengandungi isi utama penulisan wacana berita. Dua ayat pertama dalam wacana berita merupakan calon pengecaman ayat terpenting yang terbaik berdasarkan ciri kedua yang dikenal pasti dalam analisis korpus wacana berita. Fasa ini dibahagikan kepada tiga proses iaitu padanan kata signifikan dalam ayat, pengiraan nilai pemberat ayat dan pengecaman nombor ayat terpenting seperti Rajah 6.



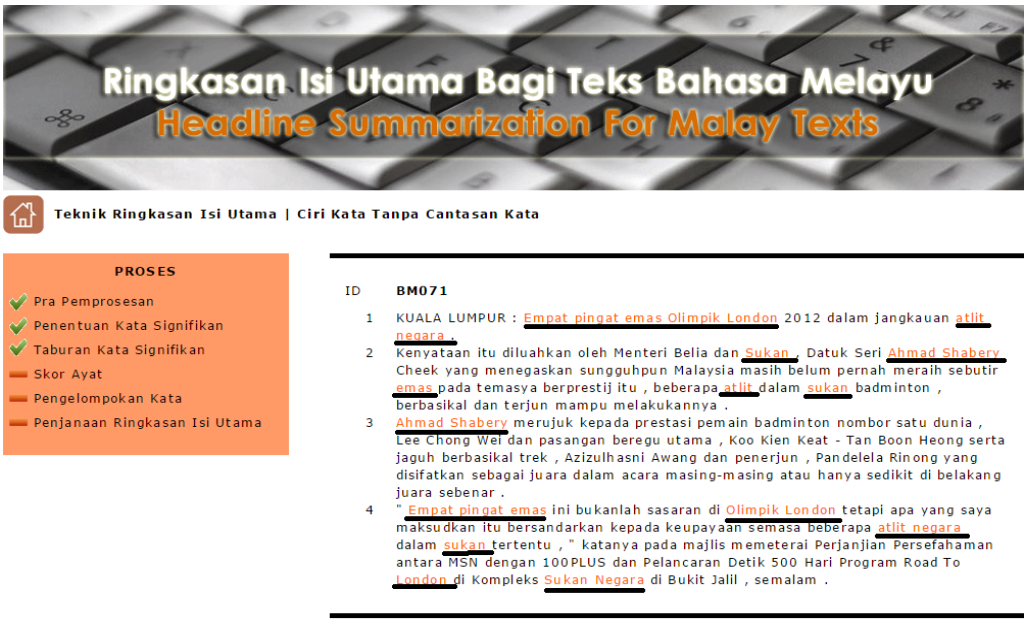
RAJAH 6. Fasa pengecaman nombor ayat terpenting

Pengiraan nilai pemberat ayat dibuat berdasarkan kebarangkalian kewujudan kata signifikan dalam ayat seperti Persamaan 4.

$$ws = \frac{ts}{m} \quad (4)$$

iaitu,
ts = bilangan kata signifikan dalam ayat
m = jumlah kata dalam ayat

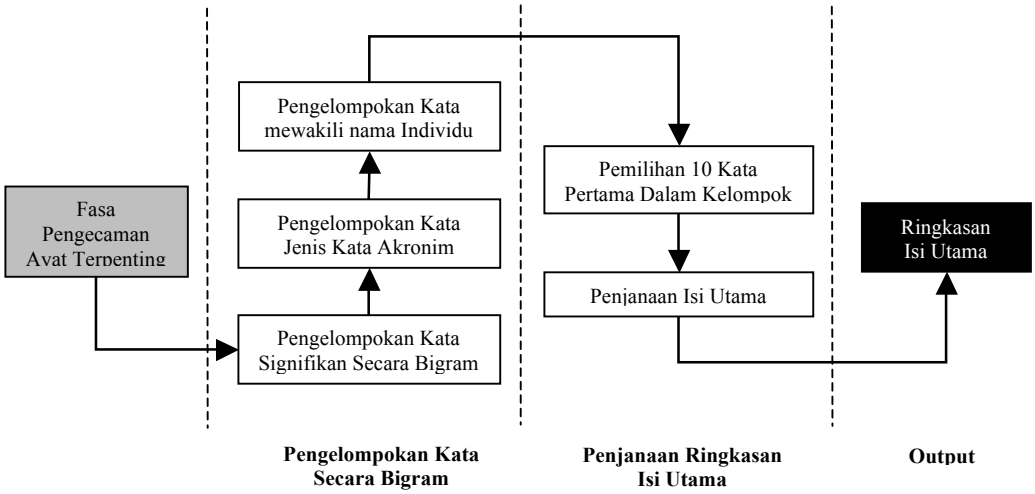
Contoh padanan kata signifikan dalam wacana berita adalah seperti Rajah 7 bagi dokumen ID BM071. Kata yang berwarna (bergaris) merupakan kata signifikan bagi dokumen tersebut. Rajah 8 pula menunjukkan contoh perbandingan nilai pemberat ayat bagi dua ayat pertama iaitu 0.7000 (Ayat 1) dan 0.1622 (Ayat 2). Ayat dengan nilai pemberat ayat tertinggi dianggap sebagai ayat yang mengandungi isi utama penulisan berita.



RAJAH 7. Padanan kata signifikan dalam wacana berita

RAJAH 8. Contoh perbandingan nilai pemberat bagi dua ayat pertama

Tujuan fasa penjanaan ringkasan isi utama adalah untuk menjanakan ringkasan isi utama secara ekstraktif berdasarkan nombor ayat terpenting yang dikenal pasti dalam fasa sebelumnya. Fasa ini dibahagikan kepada dua proses iaitu pengelompokan kata secara bigram dan penjanaan ringkasan isi utama seperti Rajah 9.



RAJAH 9. Fasa penjanaan ringkasan isi utama

Proses pengelompokan kata dalam kajian ini melibatkan tiga jenis kata iaitu (i) kata signifikan, (ii) kata berjenis kata akronim, dan (iii) kata mewakili nama individu. Proses ini dilakukan pada kata dalam ayat terpenting yang ditentukan dalam proses sebelum ini. Pengelompokan kata signifikan dibuat dengan memadankan senarai kata signifikan yang ditentukan dengan kata dalam ayat terpenting secara padanan aksara. Kata dalam ayat terpenting yang sepadan ditandakan sebagai kata signifikan. Seterusnya pengelompokan kata secara bigram dilakukan dengan mengelompokkan kata sebelum dan selepas kata yang bertanda kata signifikan sebagai satu kelompok. Pengelompokan kata berjenis kata akronim dan kata mewakili nama individu dilakukan dengan membandingkan kata dalam ayat terpenting dengan koleksi kata berjenis kata akronim dan koleksi kata mewakili nama individu. Kata dalam ayat terpenting yang sepadan dengan kata dalam koleksi kata akronim ditandakan sebagai kata akronim manakala kata dalam ayat terpenting yang sepadan dengan kata dalam koleksi kata mewakili nama individu ditandakan sebagai kata nama individu. Seterusnya pengelompokan kata secara bigram dilakukan pada kata bertanda kata akronim dan kata nama individu dengan mengelompokkan kata sebelum dan selepas kata tersebut sebagai satu kelompok.

Rajah 10 menunjukkan contoh pengelompokan kata secara bigram ke atas kata signifikan, kata akronim dan kata mewakili nama individu yang telah dikenal pasti dalam ayat terpenting diwakili dalam bentuk teks berwarna. Kata yang tidak termasuk dalam kelompok kata secara bigram dalam ayat akan digugurkan. Manakala Rajah 11 menunjukkan contoh hasil penjanaan kata ringkasan isi utama. Maksimum 10 kata pertama dipilih bagi membentuk satu ayat tunggal

RAJAH 10. Pengelompokan kata secara bigram

RAJAH 11. Hasil penjanaan ringkasan isi utama

Dua jenis penilaian intrinsik digunakan dalam kajian ini iaitu (i) kejituan, dapatan semula & skor-F dan (ii) padanan n-gram (ROUGE).

ANALISIS DATA

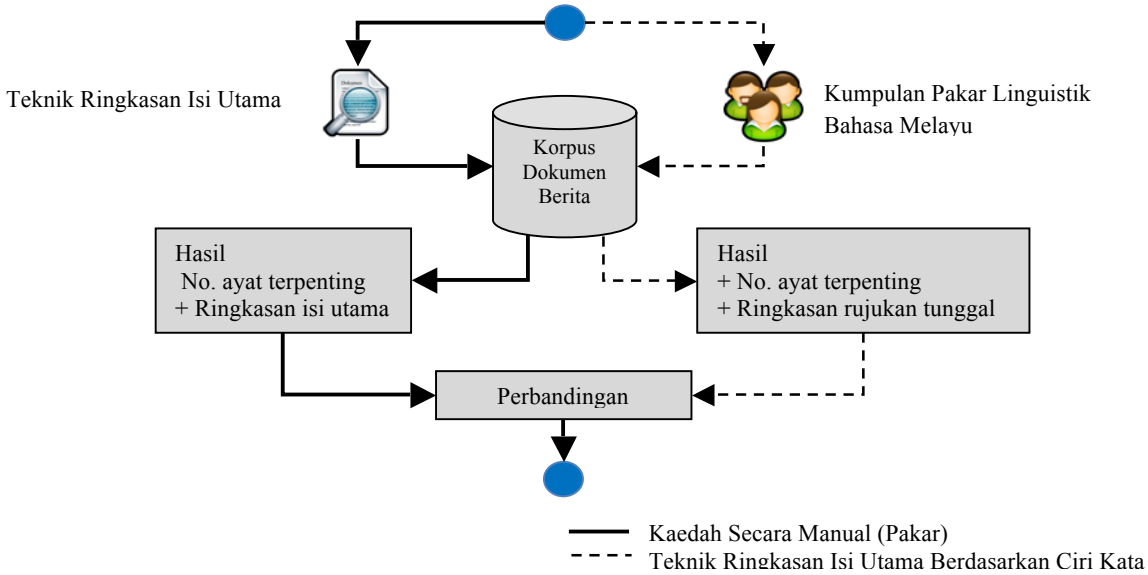
Analisis korpus wacana berita Bahasa Melayu dibahagikan kepada analisis struktur binaan wacana berita dan analisis aras permukaan teks. Analisis struktur binaan wacana berita melibatkan dua aspek iaitu struktur binaan wacana berita dan gaya penulisan wacana berita. Analisis aras permukaan teks pula melibatkan empat aspek iaitu frekuensi perkataan, ragam ayat, ayat di bahagian permulaan wacana berita dan jenis perkataan. Hasil analisis ini diringkaskan dalam Jadual 1.

JADUAL 1. Ringkasan hasil analisis korpus wacana berita Bahasa Melayu

Analisis	Aspek	Mempengaruhi Kedudukan Isi Utama Penulisan
Struktur wacana berita	Struktur binaan	Ya
	Gaya penulisan	Tidak
Aras permukaan teks	Frekuensi perkataan	Ya
	Ragam ayat	Ya
	Lokasi kedudukan perkataan	Ya
	Jenis perkataan	Ya

Berdasarkan Jadual 1, lima daripada enam aspek yang dianalisis menunjukkan aspek tersebut mempengaruhi kedudukan isi utama penulisan wacana berita Bahasa Melayu secara signifikan. Hanya aspek gaya penulisan didapati tidak signifikan kerana aspek ini lebih tertumpu kepada format penulisan wacana berita yang menjadi identiti kepada sesebuah agensi penerbitan wacana berita. Gaya penulisan juga didapati konsisten mengikut agensi penerbitannya kerana agensi cuba memastikan pembaca lebih selesa, memahami kandungan dan memenuhi jangkauannya (Rahman, 2009).

Dalam penilaian ini, perbandingan antara ringkasan isi utama yang dijana oleh teknik berdasarkan ciri kata dan ringkasan rujukan tunggal yang dihasilkan oleh kumpulan pakar dibuat dan dinilai daripada perspektif pemilihan kandungan (Steinberger & Jezek, 2009). Dua perspektif pemilihan kandungan yang digunakan adalah kesamaan kandungan (*co-selection*) dan berasaskan kandungan (*content-based*). Rajah 12 menunjukkan metodologi penilaian kualiti ringkasan isi utama bagi kajian ini.

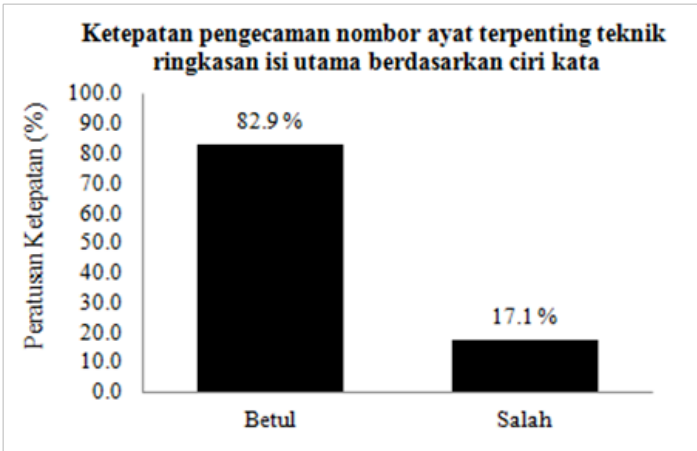


RAJAH 12. Metodologi penilaian kualiti ringkasan isi utama

Keputusan penilaian ketepatan pengecaman nombor ayat terpenting teknik ini berbandingan dengan nombor ayat terpenting yang dicam oleh kumpulan pakar lingustik Bahasa Melayu adalah seperti dalam Jadual 2 dan Rajah 13.

JADUAL 2. Ringkasan ketepatan pengecaman nombor ayat terpenting teknik ringkasan isi utama berdasarkan ciri kata

Teknik	Betul	Salah	Jumlah
Teknik Ringkasan Isi Utama Berdasarkan Ciri Kata	116	24	140



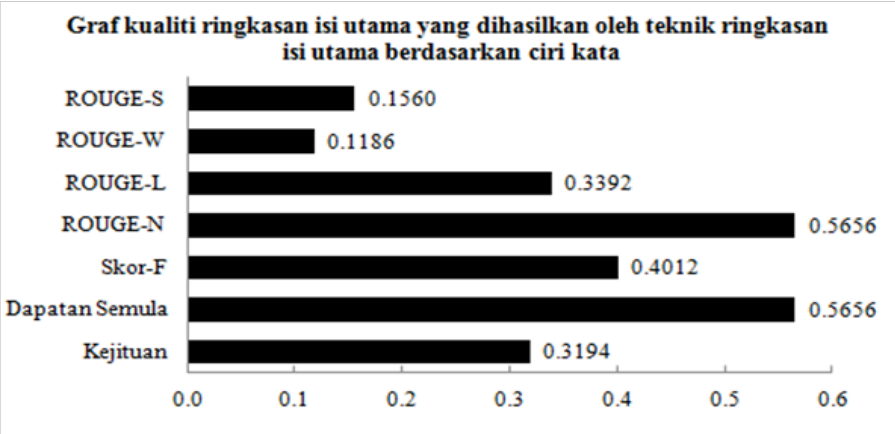
RAJAH 13. Peratus ketepatan pengecaman nombor ayat terpenting

Kualiti ringkasan isi utama yang dijana oleh teknik yang dibangunkan berbanding dengan ringkasan rujukan tunggal ditunjukkan dalam Jadual 3 dan Rajah 14.

JADUAL 3. Kualiti ringkasan isi utama yang dijanakan oleh teknik

Pemilihan Kandungan	Ukuran Metrik	
Kesamaan Pilihan (co-selection)	Kejituan (K)	0.3194
	Dapatan Semula (D)	0.5656
	Skor F (F1)	0.4012
Berasaskan Kandungan (content based)	ROUGE – N	0.5656
	ROUGE – L	0.3392
	ROUGE – W	0.1186
	ROUGE – S	0.1560

Min bagi 140 ringkasan isi utama



RAJAH 14. Graf kualiti ringkasan isi utama yang dijanakan oleh teknik

PERBINCANGAN

Keputusan pengecaman nombor ayat terpenting dalam Rajah 13 menunjukkan, ketepatan pengecaman nombor ayat terpenting oleh teknik ringkasan isi utama yang dibangunkan adalah 82.9%. Ini menunjukkan teknik ini berjaya menentukan ayat terpenting melebihi 80% berdasarkan kepada tanda aras pakar. Daripada keputusan ini, dua kesimpulan dapat dibuat iaitu kaedah penentuan kata signifikan dengan menggabungkan antara frekuensi kata dan

lokasi kedudukan kata dalam ayat berupaya menentukan kata yang signifikan dengan isi utama penulisan wacana berita Bahasa Melayu melebihi 80% dan dua ayat pertama dalam wacana berita Bahasa Melayu adalah calon ayat terpenting yang terbaik.

Keputusan penilaian kualiti dalam Jadual 3 menunjukkan, min dapatan semula (D) didapati lebih tinggi berbanding dengan min kejituan (K) dalam pemilihan kandungan berdasarkan kesamaan pilihan iaitu 0.5656 berbanding 0.3194. Min ROUGE–N (0.5656) pula didapati lebih tinggi berbanding dengan ROUGE–L (0.3392), ROUGE–W (0.1186) dan ROUGE–S (0.1560) dalam pemilihan kandungan berasaskan kandungan. Merujuk kepada jadual 3, min ROUGE–N didapati lebih tinggi berbanding dengan ROUGE–L dan min ROUGE–L didapati lebih tinggi berbanding dengan ROUGE–W. Dapatan ini adalah selaras dengan konsep penilaian ROUGE yang diperkenalkan oleh Lin (2003) berasaskan dapatan semula iaitu ukuran ROUGE–N lebih tinggi atau sama dengan ukuran ROUGE–L, ukuran ROUGE–L lebih tinggi atau sama dengan ukuran ROUGE–W dan ukuran secara n-gram (ROUGE–N) lebih tinggi atau sama dengan ukuran secara bigram (ROUGE–S) bagi setiap perbandingan dua hasil ringkasan. Daripada keputusan ini, dapat disimpulkan pertimbangan tiga ciri kata dalam fasa pengelompokan kata yang dikenal pasti dalam analisis korpus wacana berita iaitu kata signifikan, kata akronim, dan kata mewakili nama individu berjaya menghasilkan ringkasan isi utama yang lebih sempurna dan bermakna daripada aspek bahasa dengan min ROUGE–L iaitu 0.3392.

Secara keseluruhannya, didapati pertimbangan empat ciri kata iaitu lokasi kedudukan kata dalam ayat, kedudukan dua ayat pertama wacana berita, kata berjenis akronim dan kata mewakili nama individu mampu memberikan peratusan min ketepatan pengecaman ayat terpenting melebihi 80% dan kualiti ringkasan isi utama masing-masing min dapatan semula (0.5656), kejituan (0.3194), ROUGE–N (0.5656), ROUGE–L (0.3392), ROUGE–W (0.1186) dan ROUGE–S (0.1560). Pertimbangan faktor bahasa yang membina wacana berita dalam pembangunan teknik ringkasan isi utama kajian ini dilihat mampu memberi hasil yang signifikan.

KESIMPULAN

Kajian ini menjelaskan kaedah analisis dokumen berita Bahasa Melayu yang dilakukan untuk menentukan ciri kata yang signifikan dengan isi utama penulisan dokumen berita. Hasil analisis dokumen berita dalam kajian ini menunjukkan empat ciri kata yang signifikan dengan isi utama penulisan dokumen berita telah dikenal pasti iaitu, (i) lokasi kedudukan kata dalam ayat, (ii) lokasi kedudukan kata dalam dua ayat pertama, (iii) kata berjenis kata akronim, dan (iv) kata mewakili nama individu. Empat ciri kata ini dipertimbangkan dalam pembangunan teknik ringkasan isi utama bagi menghasilkan ringkasan yang berkualiti daripada aspek bahasa.

Penilaian dalam kajian ini dibuat daripada aspek kualiti ringkasan yang dihasilkan. Berdasarkan keputusan penilaian menunjukkan ketepatan pengecaman nombor ayat terpenting teknik ini adalah 82.9% dan kualiti ringkasan isi utama yang dihasilkan oleh teknik ringkasan isi utama yang dibangunkan berbanding ringkasan rujukan tunggal masing–masing iaitu kejituan (0.3194), dapatan semula (0.5656), skor–F (0.4012), ROUGE–N(0.5656), ROUGE–L(0.3392), ROUGE–W(0.1186) dan ROUGE–S(0.1560). Pertimbangan faktor bahasa dalam pembangunan teknik ringkasan isi utama mampu menghasilkan ringkasan yang berkualiti daripada aspek bahasa dan darjah ketepatan yang lebih baik. Kajian ini mampu memberi sumbangan secara tidak langsung kepada kewujudan korpus dokumen berita Bahasa Melayu khusus kepada kajian pembangunan teknik ringkasan isi utama. Antara perluasan kajian yang dicadangkan ialah mempertimbangkan aspek penyelesaian anafora (Noorhuzaimi Karimah et al., 2012) supaya ringkasan menjadi lebih tepat dan berkualiti.

Ketepatan ringkasan teks isi utama boleh ditingkatkan dengan penggunaan penyelesaian anafora yang mana ketepatan isi diperoleh melalui penyelarasan gaya bahasa dan pemilihan perkataan yang lebih tepat.

RUJUKAN

Alireza B. & Moses S. (2013). Headlines in Newspaper Editorials: A Contrastive Study. *SAGE Open. Vol. April-June(2013)*, 1-10.

Alotaiby, F. A. (2011). Automatic headline generation using character cross-correlation. Proceedings of the Association for Computational Linguistics–Human Language Technology (ACL–HLT 2011) Student Session: 117–121.

Alguliev, R. M., Aliguliyev, R. M. & Mehdiyev, C. A. (2011). Sentence Selection for Generic Document Summarization Using an Adaptive Differential Evolution Algorithm. *Swarm and Evolutionary Computation. Vol. 1*(4), 213–222.

Atkinson, J. & Munoz, Ricardo. (2013). Rhetorics-based Multi-document Summarization. *Expert System with Applications. Vol. 40*(11), 4346–4352.

Banko, M., Mittal, V. O. & Witbrock, M. J. (2000). Headline generation based on statistical translation. Proceedings of the 38th Annual Meeting on Association for Computational Linguistic (ACL–00): 318–325.

Daniel, J.A. (2008). Headline generation for Dutch newspaper articles through transformation-based learning. M.Sc Thesis: University of Groningen.

Dauzidia, F. S. & Lapalme, G. (2004). Lakhas, an arabic summarization system. Proceedings of the Document Understanding Conference 2004 (DUC 2004).

Dorr, B & Zajic, D. (2003). Cross-language Headline Generation for Hindi. *ACM Transaction on Asian Language Information Processing. Vol. 2*(3), 270-289.

Dorr, B., Zajic, D. & Schwartz, R. (2003). Hedge Trimmer: A Parse-and-Trim approach to headline generation. Proceedings of the Human Language Technology – North American Chapter of the Association for Computational Linguistics (HLT-NAACL) Workshop on Text Summarization 2003: 1–8.

El-Fishawy, N, Hamouda, A, Attiya, G. M. & Afel, M. (2014). Arabic Summarization in Twitter Social Network. *Ain Shams Engineering Journal. Vol. 5*(2), 411-420.

Edmudson, H. P. (1969). New Method in Automatic Extracting. *Journal of the Association for Computing Machinery. Vol. 16*(2), 264-285.

Foong, O. M., Oxley, A. & Sulaiman, S. (2010). Challenges and Trends of Automatic Text Summarization. *International Journal of Information and Telecommunication Technology (IJITT). Vol. 1*(1), 34-39.

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).

Gupta, V. & Lehal, G. S. (2010). A Survey of the Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence. Vol. 2*(3), 258-268.

Hamood Ali Alshalabi, Sabrina Tiun & Nazlia Omar. (2017). A Comparative Study of the Ensemble and Base Classifiers Performance in Malay Text Categorization. *Asia-Pasific Journal of Information Technology and Multimedia. Vol. 6*(2), 53-64.

Hasan, M. S. (2015). Penjanaan ringkasan isi utama berdasarkan ciri kata bagi dokumen berita Bahasa Melayu. Tesis Doktor Falsafah: Universiti Kebangsaan Malaysia.

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.

Hovy, E. & Lin, C-Y. (1997). Automated text summarization in SUMMARIST. Proceedings of the Workshop on Intelligent Scalable Text Summarization: 18-24.

Kaikhah, K. (2004). Automatic text summarization with neural networks. *Second International IEEE Conference on Intelligent System*. 40-45.

Karim, N.S, Onn, F. M., Mohammad, H. H & Mahmud, A. H . (2010). *Tatabahasa Dewan Edisi Ketiga*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Lin, J. (2009). Summarization. In. Liu, L. & Ozsu, M. (Eds.), *Encyclopedia of Database Systems*. New York: Springer.

Lin, C-Y. (2004). ROUGE: A package for automatic evaluation of summaries. Proceedings of the Association for Computational Linguistics (ACL-04) Workshop Text Summarization Branches Out : 74–81.

Lin, F-R. & Liang, C-H. (2008). Storyline-based Summarization for News Topic Retrospection. *Decision Support Systems. Vol. 45*(3), 473-490.

Lee, K. J. & Kim, J-H. (2005). Sentence compression learned by news headline for displaying in small device. Proceedings of the 2004 International Conference on Asian Information Retrieval Technology: 61–70.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstract. *IBM Journal Research and Development. Vol. 2*(2), 159-165.

Mani, I. & Maybury, M. T. (1999). *Advances in Automatic Text Summarization*. Massachusetts Avenue : Massachusetts Institute of Technology.

Muurisep, K. & Mutso, P. (2005). ESTSUM – Estonian newspaper texts summarizer. *Proceedings of the Second Baltic Conference on Human Language Technologies*: 311-316.

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.

Norshuhani Zamin & Arina Ghani. (2011). Summarizing Malay Text Documents. *World Applied Sciences Journal. Vol. 12*, 39-46.

Noorhuzaimi Karimah Mohd Noor, Shahrul Azman Noah, Mohd Juzaiddin Ab Aziz, Mohd Pouzi Hamzah. (2012). Malay Anaphor and Antecedent Candidate Identification: A Proposed Solution. Proceedings of the Asia Conference on Intelligent Information and Database (ACIIDS): 141-151

Nenkova, A. & McKeown, K. (2011). Automatic summarization. Foundations and Trends in *Information Retrieval. Vol. 5*(2-3), 103-233.

Rahman, S. N. A. (2009). Kewartawan Malaysia: Praktis & Cabaran dalam Era Revolusi Digital. Kuala Lumpur: Prentice Hall.

Shahrul Azman Mohd Noah, Nazlena Mohamad Ali & Mohd Sabri Hasan. (2018). Penentuan Fitur bagi Pengekstrakan Tajuk Berita Akhbar Bahasa Melayu (Determining Features of News Headline in Malay News Document) *GEMA Online® Journal of Language Studies. Vol.* 18(2), 154-167.

Sembok, T. M. T. (2007). *Bahasa, Kecerdasan dan Makna Sekitar Capaian Maklumat*. Bangi: Penerbitan Universiti Kebangsaan Malaysia.

Shen, D. (2009). Text summarization. In. Liu, L. & Ozsu, M. T. (Eds.). *Encyclopedia of Database Systems*. New York: Springer.

Soricut, R. & Marcu, D. (2007). Abstractive Headline Generation Using WIDL-expressions. *Information Processing and Management. Vol. 43(6),* 1536-1548.

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.

Steinberger, J. & Jezek, K. (2009). Evaluation measures for text summarization. *Computing and Informatics. Vol. 28,* 1001-1026.

Vijayapal, P., Vishnu, B., Govardhan, A. & Babu, M. Y. (2011). Statistical translation based headline generation for Telugu. *International Journal of Computer Science and Network Security. Vol. 11*(6), 295-299.

Xu, S., Yang, S. & Lau, F. C. M. (2010). Keyword extraction and headline generation using novel word features. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10): 1461-1466.

Zajic, D., Dorr, B. & Schwartz, R. (2002). Automatic headline generation for newpapers stories. Proceedings of the Document Understanding Conference (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.

Zamin, N & Ghani, A. (2011). Summarizing Malay Text Documents. World Applied Science *Journal 12 (Special Issue on Computer Applications & Knowledge Management)*: 39-46.

Zhou, L. & Hovy, E. (2003). Headline summarization at ISI. Proceedings of the Document Understanding Conference (DUC 2003).

Zhou, L. & Hovy. E. (2004). Template–filtered headline summarization. Proceedings of the Association for Computational Linguistics (ACL-04) Workshop on Text Summarization Branches Out: 56 – 60.

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. Bidang kepakaran beliau adalah *Human-Computer Interaction*.

Mohd Sabri Hasan mendapat PhD daripada Fakulti Teknologi dan Sains Maklumat, UKM pada tahun 2015. Bidang penyelidikan beliau adalah *Computational Lingsuistics* dan *Automatic Text Summarization*.